Guenther, F.H. (2002). Neural control of speech movements. To appear in: A. Meyer and N. Schiller (eds.), *Phonetics and Phonology in Language Comprehension and Production: Differences and Similarities*. Berlin: Mouton de Gruyter.

Neural control of speech movements

Frank H. Guenther

1. Introduction

Controlling speech movements for producing the syllables that make up a spoken utterance requires a complex integration of many different types of information by the brain, including auditory, tactile, proprioceptive, and muscle command representations. This chapter addresses these representations and their interactions with reference to a model of the neural processes involved in the production of speech sounds such as phonemes and syllables. The model has been developed to account for a wide variety of experimental data concerning articulator movements in adults and the development of speaking skills in children. Neural correlates of the model's components have been identified, thus allowing the model to serve as a framework for interpreting and organizing the accumulating mass of data from functional imaging studies of the human brain.

Before proceeding, it will be useful to define some reference frames that are believed to be involved in the planning of speech movements. For the present purposes, a "reference frame" can be thought of as a coordinate frame that best captures the form of information represented in a particular part of the nervous system. For example, motoneurons that project to the articulatory musculature encode information in a *muscle length reference frame*. Interactions between brain regions can be thought of as transformations of information between the corresponding reference frames. The following paragraphs define several reference frames that are important for speech production.

Muscle length reference frame. This frame describes the lengths and shortening velocities of the muscles that move the speech articulators. At the level of the facial nuclei in the brain stem, which

project to the articulatory musculature, muscle lengths or contractile states must be coded in order to position the speech articulators. However, this does not imply that the speech motor system utilizes an invariant muscle length target for each speech sound, and in fact much experimental data speak against this kind of target. For example, insertion of a bite block between the teeth forces a completely different set of muscle lengths to produce the same vowel sound, yet people are capable of compensating for bite blocks even before the first glottal pulse (Lindblom, Lubker, and Gay, 1979), illustrating the human motor system's capacity to use different muscle length configurations to produce the same phoneme under different conditions. Sensory signals from muscle spindles in the articulatory muscles also represent information about muscle lengths and shortening velocities. These signals project to the cranial nuclei and upward to primary somatosensory cortex via the ventral posterior medial nucleus (VPMN) of the thalamus.

Articulator reference frame. The *articulator reference frame*, or articulator space, refers to a reference frame whose coordinates roughly correspond to the primary movement degrees of freedom of the speech articulators (e.g., Mermelstein, 1973; Rubin, Baer, and Mermelstein, 1981; Maeda, 1990). Although it is clear that the primary movement degrees of freedom are closely related to the musculature, the articulator reference frame is often assumed to be of lower dimensionality than the muscle reference frame. For example, several muscles may move together in a synergy that effectively controls a single movement degree of freedom. Such a representation may be utilized, for example, at the level of primary motor cortex and primary somatosensory cortex. Within this view, the corticobulbar tract projections from motor cortex to facial nuclei in the brain stem perform an articulatory-to-muscular transformation, and projections from the muscle spindles to the primary somatosensory cortex via the cranial nerve nuclei and thalamus perform a muscular-to-articulatory transformation.

For the purposes of this article, the distinction between an articulator reference frame and a muscle length reference frame is

relatively unimportant, and we will therefore typically equate the two. The distinction becomes more important, however, for lower-level modeling of the kinematics and dynamics of the speech articulators (e.g., Laboissière, Ostry, and Perrier, 1995; Ostry, Gribble, and Gracco, 1996; Stone, 1991; Wilhelms-Tricarico, 1995, 1996).

Tactile reference frame. This reference frame describes the states of pressure receptors (mechanoreceptors) on the surfaces of the speech articulators, as well as the cells in primary somatosensory cortex that receive projections from pressure receptors via the cranial nerve nuclei and thalamus. For example, pressure produced when the tongue tip is pressed against the hard palate is registered by neural mechanoreceptors in the tongue and palatal surfaces. Mechanoreceptors provide important information about articulator positions when contact between articulators is made, but provide little or no information when contact is absent. Here we will use the term *orosensory* to refer to a combination of tactile and muscle length information that represents the articulator configuration accurately throughout the range of articulations used in speech.

Constriction reference frame. Several researchers have proposed reference frames for speech production whose coordinates describe the locations and degrees of key constrictions in the vocal tract (e.g., Browman and Goldstein, 1990; Coker, 1976; Guenther, 1994, 1995a; Saltzman and Munhall, 1989). Typical constrictions include a tongue body constriction, tongue tip constriction, and lip constriction. It is important to note that the relationship between the constriction frame and the articulator frame is one-to-many; that is, a given set of constriction locations and degrees can be reached by an infinite number of different articulator configurations. In the case of a vowel, for example, the same target tongue body constriction could be reached with the mandible high and the tongue body low relative to the mandible under normal conditions, or with the mandible lower and the tongue body higher if a bite block is present. This one-to-many relationship makes it possible for a movement controller that uses invariant constriction targets and an appropriate mapping

between the constriction and articulator frames to overcome constraints on the articulators (such as a bite block) by utilizing different articulator configurations to produce the same constrictions (e.g., Saltzman and Munhall, 1989; Guenther, 1992, 1994, 1995a). This ability to use different movements to reach the same goal under different conditions, called *motor equivalence*, is a ubiquitous property of biological motor systems and is addressed further in Section 4. In this chapter, we will assume that constriction information is part of the orosensory representation of the vocal tract.

Acoustic reference frame. The acoustic reference frame describes important properties of the acoustic signal produced by the vocal tract (e.g., formant frequencies, amplitudes, and bandwidths).

Auditory perceptual reference frame. The central nervous system has access to acoustic signals only after transduction by the auditory system. In the current chapter, the term "auditory perceptual" will be used to refer to the transduced version of the acoustic signal (cf. Miller, 1989; Savariaux, Perrier, and Schwartz, 1995) as represented in auditory cortical areas. Although the important aspects of the auditory representation for speech are still not fully understood, several researchers have attempted to characterize them. In the implementation of the DIVA model described below, we utilize the auditory perceptual frame proposed by Miller (1989), although we acknowledge the incompleteness of this auditory representation for capturing all of the perceptually important aspects of speech sounds. This auditory perceptual space is made up of three dimensions:

$$x_1 = \log(F1 / SF0)$$

$$x_2 = \log(F2 / F1)$$

$$x_3 = \log(F3 / F2)$$

where $F1$, $F2$, and $F3$ are the first three formants of the acoustic signal, and $SF0 = 168(F0/168)^{1/3}$, where $F0$ is the fundamental frequency of the speech waveform. This space was chosen by Miller (1989) in part because these coordinates remain relatively constant

for the same vowel when spoken by men, women, and children, unlike formant frequencies.

2. The DIVA model of speech production

Our laboratory has developed a neural network model of speech motor skill acquisition and speech production called the DIVA model (for Directions Into Velocities of Articulators). The model addresses the neural representations underlying speech production, as well as the nature of the interactions (or *mappings*) between these representations. The model describes speech production processes from the syllable level "on down"; i.e., it addresses the transformation of syllable- or phoneme-sized speech targets into the muscle commands that carry out the desired speech sound. For an account of the higher-level processes involved in transforming sentences into syllables for production, see Levelt (1989; Levelt, Roelofs, and Meyer, 1999), and for a different perspective on syllable production, see Fujimura (2000).

A simplified block diagram of the DIVA model is provided in Figure 1. Each block in the diagram corresponds to a set of neurons that together constitute a neural representation, and arrows and filled semicircles correspond to mappings between the neural representations. Three of the mappings in the model, indicated by filled semicircles in the figure, are tuned during a "babbling stage" in which random movements of the speech articulators provide tactile, proprioceptive, and auditory feedback signals. This information is used to tune parameters that correspond to synaptic weights. These synaptic weights constitute the learned neural mappings, which effectively encode speaker-specific information about the relationships between articulator movements and their tactile, proprioceptive, and auditory consequences. After learning, these mappings are used for phoneme production. Because the model is a self-organizing neural network whose parameters are tuned during an action-perception cycle, it requires no explicit knowledge about the physical geometry of the vocal tract being controlled.

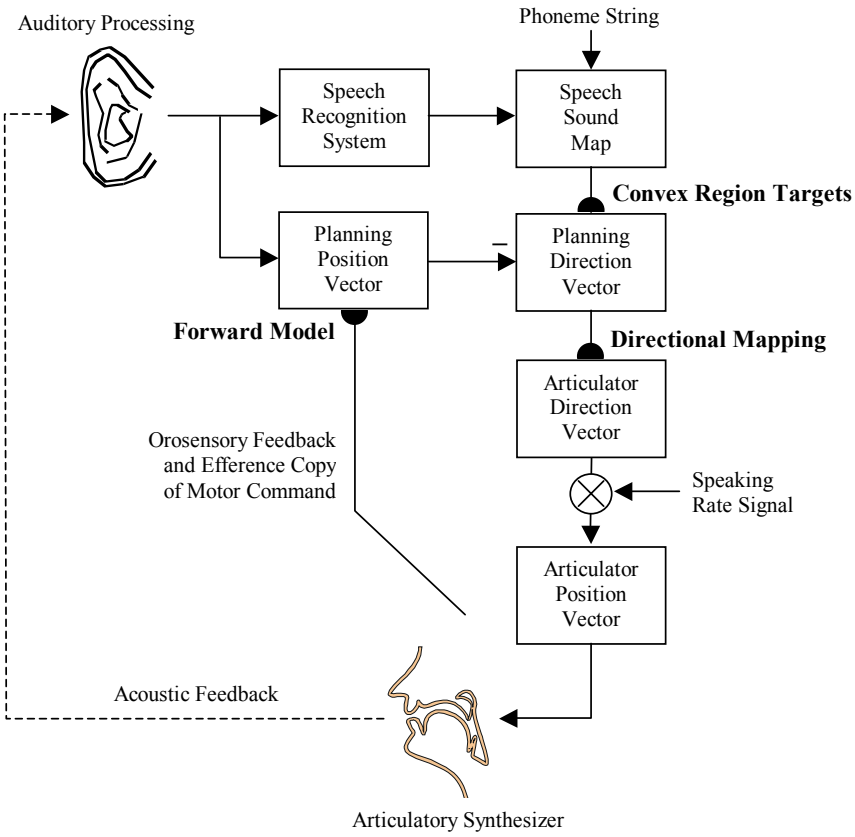

Figure 1. Overview of the DIVA model. Filled semicircles represent learned neural mappings. See text for details.

The synaptic weights of the first mapping, labeled "convex region targets" in the figure, encode targets for each phoneme the model encounters during babbling. These targets are defined in a planning space made up of auditory and orosensory dimensions. For example, the target for vowel sounds specifies a range of acceptable values of formant ratios (see Section 1). To account for the human ability to learn phoneme-specific and language-specific limits on acceptable articulatory and acoustic variability, the learned speech sound targets take the form of multidimensional regions, rather than points, in the planning space. This notion of phonemic targets as multidimensional regions provides a simple and unified explanation for many long-

studied speech phenomena (see Guenther, 1995a for details). This topic is addressed in Section 3.

The second neural mapping, labeled "directional mapping" in the figure, transforms desired movement directions in planning space into movement directions in an articulator space closely related to the vocal tract musculature. This mapping embodies a solution to the inverse kinematic problem for control of a redundant manipulator (in this case, the vocal tract). The model posits that, during babbling, the brain learns a transformation from desired movement directions in auditory and orosensory spaces into articulator velocities that carry out the desired movement directions. The use of this mapping to control the model's articulator movements is closely related to pseudoinverse-style control techniques in robotics (e.g., Liégeois, 1977), and the resulting controller is capable of automatically compensating for constraints and/or perturbations applied to the articulators (Guenther, 1994, 1995a; Guenther and Micci Barreca, 1997), thus accounting for the motor equivalent capabilities observed in humans when speaking with a bite block or lip perturbation. This topic is addressed further in Section 4.

The third mapping, labeled "forward model" in the figure, transforms orosensory feedback from the vocal tract and an efference copy of the motor outflow commands into a neural representation of the auditory signal produced by the current vocal tract configuration. This forward model allows the system to control speech movements by indicating the vocal tract's position with the planning space without relying on auditory feedback, which may be absent or too slow for use in controlling ongoing articulator movements.

According to the model, the production of a speech sound takes place as follows. First, a cell corresponding to the sound in the speech sound map of Figure 1 is activated. This has the effect of reading out that sound's target to the planning direction vector stage of the model. Cells here represent the difference between the target and the current position of the vocal tract as represented in the planning space. This difference defines the desired movement direction in the planning space, which consists of auditory and

orosensory dimensions. The desired movement direction in planning space is transformed into a commanded movement direction in articulator space via the directional mapping projecting from the planning direction vector to the articulator direction vector stages. These directional commands are translated into positional commands at the articulator position vector stage. As the vocal tract moves to the target, the planning position vector is continuously updated via orosensory feedback and an efference copy of the motor command; this information is mapped into the planning space via the forward model.

Computer simulations have been used to verify that the model provides a unified explanation for a wide range of data on articulator kinematics and motor skill development (Guenther, 1994, 1995a,b; Guenther, Hampson, and Johnson, 1998; Callan et al., 2000) that were previously addressed individually rather than in a single model. The model's explanations for several speech production phenomena are discussed in the next two sections, which deal with two important issues addressed by the model: the nature of the brain's "targets" for speech motor control, and the manner in which the nervous system achieves motor equivalence in speech.

3. The nature of speech sound targets

Most accounts of speech production involve some sort of "target" that the motor system hopes to achieve in order to produce a particular speech sound. For example, phoneme targets in the task-dynamic model (Saltzman and Munhall, 1989) take the form of locations and degrees of key constrictions of the vocal tract. Targets in the DIVA model take the form of regions in a planning space consisting of auditory and orosensory dimensions (e.g. formant ratios and vocal tract constrictions). Each cell in the model's speech sound map (see Figure 1) represents a different sound (phoneme or syllable). The synaptic weights on the pathways projecting from a speech sound map cell to cells in the planning direction vector represent a target for the corresponding speech sound in planning

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 in turn causes learning to occur in the synaptic weights of the pathways projecting from that cell, thereby allowing the model to modify the target for the speech sound based on the current configuration of the vocal tract.

To explain how infants learn phoneme-specific and language-specific limits on acceptable articulatory variability, the targets take the form of convex regions in planning space. This "convex region theory" is a generalization of Keating's (1990) "window model" of coarticulation to a multi-dimensional movement planning space consisting of auditory and constriction dimensions in addition to articulatory dimensions (see Guenther, 1995 for further discussion of this topic).

Figure 2 schematizes the learning sequence for the vowel /i/ along two dimensions of planning space, corresponding to lip aperture and tongue body height. The first time the phoneme is produced during babbling, synaptic weights that project from the speech sound map cell for /i/ are adjusted to encode the position in planning space that led to proper production of the phoneme on this trial. In other words, the model has learned a target for /i/ that consists of a single point in the planning space, as schematized in Figure 2a. The next time the phoneme is babbled, the speech sound map cell expands its learned target to be a convex region that encompasses the previous point and the new point in planning space, as shown in Figure 2b; this can occur via a simple and biologically plausible learning law (Guenther, 1995a). In this way, the model is constantly expanding its convex region target for /i/ to encompass all of the various vocal tract configurations that can be used to produce /i/.

An important aspect of this work concerns how the nervous system extracts the appropriate forms of auditory and orosensory information that define the different speech sounds. For example, 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 not only vowels but also velar, alveolar, and dental stops? 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?

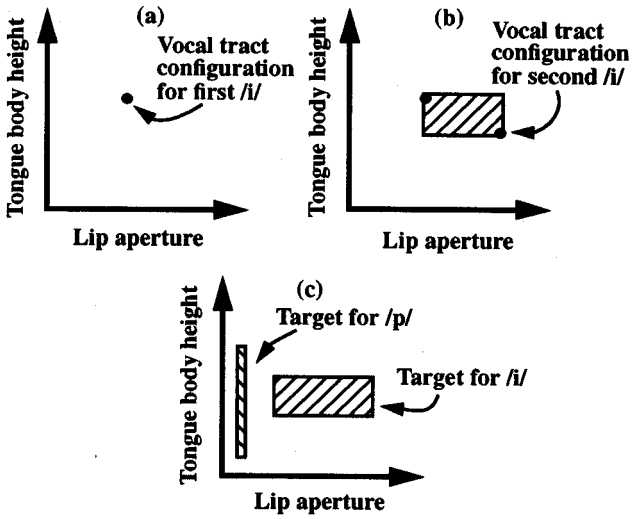

Figure 2. Learning of the convex region target for the vowel /i/ along planning dimensions corresponding to lip aperture and tongue body height. (a) The first time /i/ is produced during babbling, the learned target is simply the configuration of the vocal tract when the sound was produced. (b) The second time /i/ is babbled, the convex region target is expanded to encompass both vocal tract configurations used to produce the sound. (c) Schematized 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/.

The manner in which targets are learned in the DIVA model provides a unified answer to these questions. Consider the convex regions that result after many instances of producing the vowel /i/ and the bilabial stop /p/ (Figure 2c). The convex region for /p/ does

not vary over the dimension of lip aperture but varies largely over the dimension of tongue body height; this is because all bilabial stops that the model has produced have the same lip aperture (corresponding to full closure of the lips), but tongue body height has varied. In other words, the model has learned that lip aperture is the important dimension for producing the bilabial stop /p/. Furthermore, whereas lip aperture is the important 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 Figure 2c. Finally, since convex region learning relies on language-specific recognition of phonemes by the infant, the shapes of the resulting convex regions will vary from language to language.

As currently implemented, the model implicitly assumes that an infant is able to properly perceive a speech sound before he/she can learn to produce the sound properly. Furthermore, it is assumed that the infant can identify individual phonemes within a syllable. These assumptions are made to simplify the learning process in computer simulations of the model and are not being posed as hypotheses concerning speech development in infants. Although we believe the model is general enough to accommodate several different possibilities regarding the size of the units learned by infants (e.g., syllables vs. phonemes) and the relationship between perceptual and production learning, these complex issues are currently beyond the scope of the model's explanatory capabilities.

An interesting property of the model's learning process is that the model can learn to "ignore" totally unimportant orosensory or auditory dimensions by allowing variability throughout the entire range of such dimensions. For example, no harm is done by including dimensions that are important only for some languages but not for others, since speakers of languages that do not use a dimension can simply learn to ignore it. The babbling process causes the system to learn small target ranges for acoustically important planning dimensions (i.e., those that must be carefully controlled to successfully produce the desired sound, such as formant ratios for a vowel), and large ranges for relatively unimportant dimensions. When moving to a learned target, the model moves to the point on

the convex region target that is closest to the current configuration of the vocal tract. If the vocal tract configuration is already within the range for a particular target dimension, no further movement is planned along this dimension. The effect of these properties on articulator movements is a general tendency not to move an articulator unless it needs to be moved, thus allowing the model to make very efficient movements (see Guenther, 1995a; Guenther and Micci Barreca, 1997; Guenther, Hampson, and Johnson, 1998; Perkell et al., 2000).

The convex region theory of the targets of speech provides a unified explanation for a number of long-studied speech production phenomena. A brief summary of some of these data explanations is provided below; see Guenther (1995a) for further detail.

Convex region targets provide a natural framework for interpreting data on motor variability in speech: the motor system is careful to control movements along dimensions that are important for a sound (i.e., dimensions with small target ranges), but not movements along dimensions that are not important (those with large target ranges). The model accordingly shows more variability for acoustically unimportant dimensions as compared to acoustically important dimensions, as seen in the experimental results of Perkell and Nelson (1985).

The theory's explanation for carryover coarticulation is simple and straightforward: when producing a phoneme from different initial configurations of the vocal tract, different positions on the convex region target will be reached, as schematized in Figure 3, since the model moves to the closest point on the target region. The end effect of this is that the configuration used to produce a sound will depend on which sound precedes it, with the model choosing a configuration that is as close as possible to the preceding configuration. In contrast to the view of carryover coarticulation as the result of mechano-inertial effects, carryover coarticulation in the DIVA model is "planned" in the sense that it results from explicit movement commands. This planning does not require advance knowledge of later segments, but instead arises from the interaction between the configuration of the vocal tract at the start of a segment

and the convex region target for the segment. As pointed out by Daniloff and Hammarberg (1973), the mechano-inertial explanation is inadequate since large carryover effects are seen at low speeds and may spread over two or three segments, indicating a deliberate process for producing these effects. Based on a study requiring subjects to begin an utterance before knowing its end, Whalen (1990) also hypothesized that carryover effects are probably largely planned, but to a lesser degree than anticipatory effects.

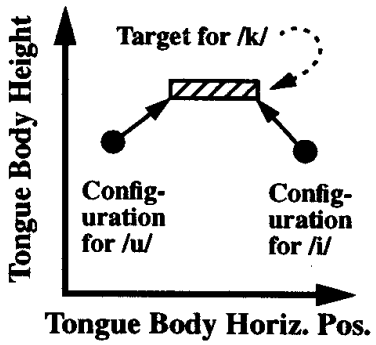

Figure 3. Convex region theory account of carryover coarticulation in /k/ production. Approaching the target for /k/ from the configuration corresponding to the back vowel /u/ in "luke" leads to a final tongue body configuration that is further back than when approaching from the configuration corresponding to the front vowel /i/ in "leak".

The convex region theory's explanation of anticipatory coarticulation posits that the target region for a speech sound is reduced in size based on context in order to provide a more efficient sequence of articulator movements. Because the amount of anticipatory coarticulation is limited by the size of the convex region targets in the model, it accounts for experimental results showing decreased coarticulation in cases where smaller targets are necessitated, including speech in languages with more crowded vowel spaces (Manuel, 1990), speech hyperarticulated for clarity (Picheney, Durlach, and Braida, 1985, 1986; Lindblom and MacNeilage, 1986) or stress (De Jong, Beckman, and Edwards, 1993), and speech of small children who may have not yet learned

the full range of variation allowed for some phonemes (Thompson and Hixon, 1979; Kent, 1983; Sereno and Lieberman, 1987).

The model also provides an explanation for data regarding the effects of speaking rate on articulator movements (Guenther, 1995a). Shrinking of target regions for better accuracy during slower speech, as suggested by the well-known speed-accuracy trade-off known as Fitts' Law (e.g., Woodworth, 1899; Fitts, 1954), leads to differential effects for vowels and consonants: the speed of consonant movements decreases as one would expect, but the speed of vowel movements remains approximately constant or even increases. This is in concert with experimental data on speaking rate effects (e.g., Gay, Ushijima, Hirose, and Cooper, 1974). The model shows how a single control process can produce these differential effects due to inherent differences in the shapes of the target convex regions for vowels and consonants. Despite the differential effects on movement velocities, the ratio of maximum velocity to movement distance increases by about the same amount for the two sound types, again as seen in human speaking data. Furthermore, cross-speaker differences in strategies for increasing speaking rate are captured by variation of a single parameter in the model.

4. Motor equivalence and directional mappings

Motor equivalence is the ability to carry out the same task using different motor means. For example, people are capable of producing written letters with very similar shapes using their wrist and fingers or shoulder and elbow (Merton, 1972), their dominant or non-dominant arms (Raibert, 1977; Wright, 1990), and even using pens attached to their feet or held in their teeth (Raibert, 1977). Motor equivalence is seen in a wide variety of human behaviors, including handwriting, reaching (e.g., Cruse, Brüwer, and Dean, 1993), and speaking (e.g., Abbs and Gracco, 1984; Lindblom, Lubker, and Gay, 1979; Savariaux, Perrier, and Orliaguet, 1995), and in a wide variety of species, including turtles (Stein, Mortin, and Robertson, 1986) and frogs (Berkinblit, Gelfand, and Feldman, 1986). The ubiquity of motor equivalence is no doubt the evolutionary result of its utility:

animals capable of using different motor means to carry out a task under different environmental conditions have a tremendous advantage over those that cannot.

An enlightening example of motor equivalent behavior is the ability to use redundant degrees of freedom to compensate for temporary constraints on the effectors while producing movement trajectories to targets. For example, people normally use jaw movements during speech, but they can also successfully produce phonemes with a bite block clenched in their teeth by increasing lip and tongue movements to compensate for the fixed jaw. Compensation occurs immediately and automatically; i.e., without requiring practice with the bite block (Lindblom, Lubker, and Gay, 1979), though a smaller additional increment in performance can be gained with some practice (McFarland and Baum, 1995; Baum, McFarland, and Diab, 1996).

The DIVA model has been formulated to deal with the problem of motor equivalence. The model stresses *automatic* compensation, i.e.:

- it successfully compensates for constraints on the effectors even if the constraints have never before been experienced,
- it does not require new learning or practice under the constraining conditions, and
- it does not invoke special control strategies to deal with constraints, instead utilizing the same control scheme used during unconstrained movements.

Automatic compensation can greatly reduce the computational requirements of movement planning, potentially freeing up cognitive resources for more important or more difficult tasks.

In order to understand the motor equivalent capabilities of the model, it is useful to consider a simplified view of the movement control process wherein movement trajectories are planned within some reference frame (the planning frame), and these trajectories are mapped into a second reference frame that relates closely to the effector or articulator system that carries out the movements. For example, one can consider speech production as the process of formulating a trajectory within a planning frame to pass through a sequence of targets, each corresponding to a different phoneme in the

string being produced. The dimensions of this planning frame might correspond to acoustic quantities or locations and degrees of key constrictions in the vocal tract. The planned trajectory can then be mapped into a set of articulator movements that realize the trajectory. The articulator movements are defined within an articulator reference frame that relates closely to the musculature or primary movement degrees of freedom of the speech articulators. The process of mapping from the planning frame to the articulator frame need not wait until the entire trajectory has been planned, but instead may be carried out in concurrence with trajectory planning.

Based on a number of theoretical analyses and numerical simulation results, we have posited that maximal automatic compensation is possible if trajectory planning is carried out in a reference frame that relates closely to the task space for the movement (e.g., 3D space for reaching or an acoustic-like space for speaking), rather than a frame that relates more closely to the effector or articulator system (Guenther, 1992, 1994, 1995a,b; Bullock, Grossberg and Guenther, 1993; Guenther and Micci Barreca, 1997; Guenther, Hampson, and Johnson, 1998). The use of auditory and orosensory dimensions that relate closely to the acoustic signal in the model's planning space is motivated by these findings.

Trajectories planned in task space must still be carried out by articulator or effector movements. One possibility is to use a position-to-position mapping from task space to articulator space; e.g., each point in acoustic space could be mapped to an articulator configuration that would achieve that acoustic result. Another possibility is to use a directional mapping from desired movement directions in task space into movement directions in articulator space. The DIVA model uses the latter form of mapping because it provides the automatic compensation for externally imposed constraints on effector motion (Guenther, 1992, 1994, 1995a,b; Guenther, Hampson, and Johnson, 1998). The use of a directional mapping for movement control is closely related to robotic controllers that utilize a generalized inverse, or pseudoinverse, of the Jacobian matrix relating task and effector spaces (e.g., Baillieul, Hollerbach, and

Brockett, 1984; Hollerbach and Suh, 1985; Klein and Huang, 1983; Liégeois, 1977; Mussa-Ivaldi and Hogan, 1991; Whitney, 1969).

The ability to reach targets in pseudoinverse-style controllers such as the DIVA model is very robust to error in the directional mapping. This can been seen in the following example. Imagine an intended straight-line movement to a target in task space (e.g., auditory space for a speech sound), as schematized in Figure 4. Assume that a 30^o error in the directional mapping causes the actual trajectory to veer upward from the desired straight-line trajectory. The planning direction vector (indicated by dashed arrows in the figure) always points from the current position to the target. As the actual trajectory moves further away from the desired trajectory, the planning direction vector points more and more downward to counteract the error. The system thus "steers in" toward the target. As long as the directional mapping is off by less than 90^o, the target will be successfully reached, although for large directional errors the trajectory will deviate significantly from a straight line.

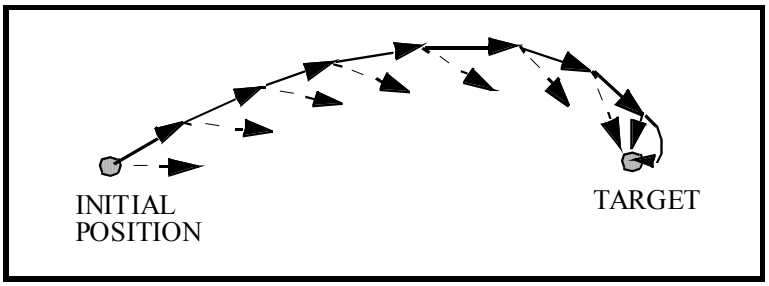

Figure 4. Robustness to error in the directional mapping for targeted movements. Here a 30^o error in the mapping causes the actual trajectory to veer from the desired straight-line trajectory. Dashed arrows indicate the desired task space movement direction at each point along the trajectory. As the actual trajectory moves further away from the desired trajectory, the task space direction vector points more and more downward to counteract this error in movement direction, allowing the system to "steer in" toward the target. As long as the directional mapping is off by less than 90^o, the target will be successfully reached.

This automatic error-correction property has important implications for biological movement control. First, it suggests how a person can easily overcome constraints on the effectors (such as a

cast limiting arm movement during reaching or a bite block limiting jaw movement during speaking) that effectively introduce error in the directional mapping, and thus provides an explanation for one form of motor equivalence. Simulations verifying the abilities of the DIVA model to overcome errors in the directional mapping due to blockage of one or more speech articulators are provided elsewhere (Guenther, 1992, 1994, 1995a,b; Guenther, Hampson, and Johnson, 1998). Second, it implies that even coarsely learned directional mappings, such as those possessed by an infant in the early months of life, can be used to reach objects or produce speech sounds, although with imperfect movement trajectories. Finally, it shows how error correction capabilities can automatically arise from the same mechanism used to control normal movements, unlike a controller that aims for postural targets and must somehow choose a new postural target if the normal target is inaccurate or unreachable due to external constraints.

5. Hypothesized neural correlates of the DIVA model

One advantage of the neural network approach is that it allows one to analyze the brain regions involved in speech in terms of a well-defined theoretical framework, thus allowing a deeper understanding of the brain mechanisms underlying speech. Figure 5 illustrates hypothesized neural correlates for several central components of the DIVA model. These hypotheses are based on a number of neuroanatomical and neurophysiological studies, including lesion/aphasia studies, brain imaging studies involving magnetoencephalography (MEG), positron emission tomography (PET), and functional magnetic resonance imaging (fMRI), and single-cell recordings from cortical and subcortical areas in animals. (For a related review of neuroimaging data on speech, see Indefrey and Levelt, 2000.)

The pathway labeled ‘a’ in the figure corresponds to projections from premotor cortex to primary cortex, hypothesized to underlie feedforward control of the speech articulators. Pathway b represents hypothesized projections from premotor cortex (lateral BA 6) to

higher-order auditory cortical areas in the superior temporal gyrus (BA 22) and orosensory association areas in the supramarginal gyrus (BA 40). These projections are hypothesized to carry target sensations associated with motor plans in premotor cortex. For example, premotor cortex cells representing the syllable /bi/ project to higher-order auditory cortex cells; these projections represent an expected sound pattern (i.e., the auditory representation of the speaker's own voice while producing /bi/). Similarly, projections from premotor cortex to orosensory areas in the supramarginal gyrus represent the expected pattern of somatosensory stimulation during /bi/ production. Pathway b is hypothesized to encode the convex region targets for speech sounds in the DIVA model, corresponding to the pathway between the speech sound map and planning direction vector in Figure 1.

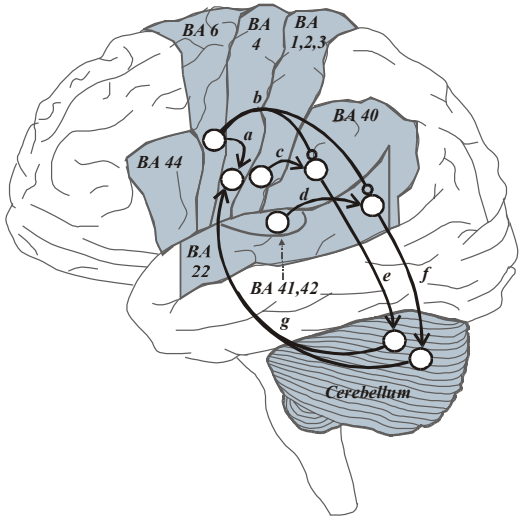

Figure 5. Hypothesized neural correlates of several central components of the DIVA model. BA = Brodmann's Area. See text for details.

One interesting aspect of the model in Figure 5 is the role of auditory cortical areas in speech production as well as speech perception. According to the model, auditory "targets" project from premotor cortical areas to the posterior superior temporal gyrus (pathway b), where they are compared to incoming auditory

information from primary auditory cortex (pathway d). The difference between the target and the actual auditory signal represents an "error" signal that is mapped through the cerebellum (pathway f), which transforms the auditory error into a motor velocity signal that can act to zero this error (pathway g). This projection through the cerebellum to motor cortex forms a component of the Directions Into Velocities of Articulators mapping that gives the DIVA model its name.

Evidence that auditory cortical areas in the superior temporal gyrus and temporal plane are involved in speech production comes from a number of neuroimaging studies. For example, Hickok et al. (2000) report activation in left posterior superior temporal gyrus areas (planum temporale, superior temporal sulcus) during a PET visual object naming task in which the subject's auditory feedback of his/her own productions was masked with noise. Bookheimer et al. (1995) report activations near primary auditory cortex in a similar task. Paus et al. (1996) also reported activation in the area of the left planum temporale during a PET object naming task. These authors attributed this activation to "motor-to-sensory discharges", compatible with pathway b in Figure 5. This interpretation receives support from an MEG study by Levelt et al. (1998), who showed that the auditory cortical activations during speech production slightly preceded the initiation of articulatory processes. All of these results provide support for the notion of auditory perceptual targets for speech production, in keeping with a central aspect of the DIVA model (e.g., Guenther, 1995b; Guenther et al., 1998; see also Perkell et al., 1995; Bailly et al., 1991).

The model also proposes a novel role for the supramarginal gyrus (BA 40) in speech production. This brain region has been implicated in phonological processing for speech perception (e.g., Caplan, Gow, and Makris, 1995; Celsis et al., 1999), as well speech production (Geschwind, 1965; Damasio and Damasio, 1980). The current model proposes that, among other things, the supramarginal gyrus represents the difference between target oral sensations (projecting from premotor cortex via pathway b in Figure 5) and the current state

of the vocal tract (projecting from somatosensory cortex via pathway c). This difference represents the desired movement direction in orosensory coordinates and is hypothesized to map through the cerebellum to motor cortex, thus constituting a second component of the Direction Into Velocities of Articulators mapping.

Not shown in Figure 5, for the sake of clarity, is the insular cortex (BA 43), buried within the sylvian fissure. The anterior insula has been shown to play an important role in speech articulation (e.g., Dronkers, 1996). This region is contiguous with the frontal operculum, which includes portions of the premotor and motor cortices related to oral movements. We adopt the view that the anterior insula has similar functional properties to the premotor and motor cortices. This view receives support from fMRI studies showing activation of anterior insula during non-speech tongue movements (Corfield et al., 1999), PET results showing concurrent primary motor cortex and anterior insula activations during articulation (Fox et al., 2001), and PET results showing concurrent lateral premotor cortex and anterior insula activations during articulation (Wise et al., 1999).

An important purpose of the model outlined in Figure 5 is to generate predictions that serve as the basis for focused functional imaging studies of brain function during speech. For example, the model of Figure 5 predicts that perturbation of a speech articulator such as the lip during speech should cause an increase in activation in the supramarginal gyrus, since the perturbation will cause a larger mismatch between orosensory expectations and the actual orosensory feedback signal. The model further predicts that extra activation will be seen in the cerebellum and motor cortex under the perturbed condition, since pathway e in Figure 5 would transmit the extra supramarginal gyrus activation to the cerebellum and on to motor cortex (pathways e, g). We are currently testing these and other predictions of the model using fMRI and MEG.

6. Summary

This chapter has described a model of the neural processes underlying speech production. This model has been designed to provide a simple and unified account for a wide range of experimental data, including functional brain imaging, psychophysical, physiological, anatomical and acoustic data. The model has also been used to study the effects of auditory feedback on speech in normally hearing individuals, hearing impaired individuals, and cochlear implant recipients (Perkell et al., 2000). According to the model, the goals of speech movements are regions in a planning space whose dimensions relate closely to the acoustic signal. It is hypothesized that projections from premotor cortex to higher-order auditory and somatosensory cortical areas encode these sound targets. Planned trajectories are mapped into articulator movements via a directional mapping between the planning and articulator spaces. This mapping is hypothesized to involve a pathway from higher-order auditory and somatosensory cortical areas through the cerebellum to the motor cortex.

7. References

Abbs, James H., and Gracco, Vincent L. (1984). Control of complex motor gestures: Orofacial muscle responses to load perturbations of lip during speech. *Journal of Neurophysiology*, **51**, pp. 705-723.

Baillieul, John, Hollerbach, John, and Brockett, Roger W. (1984). Programming and control of kinematically redundant manipulators. *Proceedings of the 23rd IEEE Conference on Decision and Control*, pp. 768-774. New York: IEEE.

Bailly, Gerard, Laboissière, Rafael, and Schwartz, Jean-Luc (1991). Formant trajectories as audible gestures: An alternative for speech synthesis. *Journal of Phonetics*, **19**, pp. 9-23.

Baum Shari R., McFarland David H., Diab, Mai (1996). Compensation to articulatory perturbation: perceptual data. *Journal of the Acoustical Society of America*, **99**, pp. 3791-3794.

Berkinblit, Misha B., Gelfand, Israil M., and Feldman, Anatol G. (1986). A model of the aiming phase of the wiping reflex. In S. Grillner, P.S.G. Stein, D.G.

Stuart, H. Forssberg, and R.M. Herman (eds.), *Neurobiology of vertebrate locomotion* (pp. 217-227). London: Macmillan.

Bookheimer, Susan Y., Zeffiro, Thomas A., Blaxton, Teresa, Gaillard, William, and Theodore, William (1995). Regional cerebral blood flow during object naming and word reading. *Human Brain Mapping*, **3**, pp. 93-106.

Browman, Catherine, and Goldstein, Louis (1990). Gestural specification using dynamically-defined articulatory structures. *Journal of Phonetics*, **18**, pp. 299-320.

Bullock, Daniel, Grossberg, Stephen, and Guenther, Frank H. (1993). A self-organizing neural network model for redundant sensory-motor control, motor equivalence, and tool use. *Journal of Cognitive Neuroscience*, **5**, pp. 408-435.

Callan, Daniel, Kent, Raymond, Guenther, Frank H., and Vorperian, Houri K. (2000). An auditory-feedback-based neural network model of speech production that is robust to developmental changes in the size and shape of the articulatory system. *Journal of Speech, Language and Hearing Research*, **43**, 721-736.

Caplan, David, Gow, David, and Makris, Nikos (1995). Analysis of lesions by MRI in stroke patients with acoustic-phonetic processing deficits. *Neurology*, **45**, pp. 293-298.

Celsis, Pierre, Boulanouar, Kader, Ranjeva, J.P., Berry, Isabelle, Nespoulous, Jean-Luc, and Chollet, F. (1999). Differential fMRI responses in the left posterior superior temporal gyrus and left supramarginal gyrus to habituation and change detection in syllables and tones. *NeuroImage*, **9**, pp. 135-144.

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

Corfield, Douglas R., Murphy, Kevin, Josephs, O., Fink, Gereon R., Frackowiak, Richard S.J., Guz, Abraham, Adams, Lewis, and Turner, R. (1999). Cortical and subcortical control of tongue movement in humans: A functional neuroimaging study using fMRI. *Journal of Applied Physiology*, **86**, pp. 1468-1477.

Cruse, Holk, Brüwer, M., and Dean, Jeffrey (1993). Control of three- and four-joint arm movement: Strategies for a manipulator with redundant degrees of freedom. *Journal of Motor Behavior*, **25**(3), pp. 131-139.

Damasio, Hanna, and Damasio, Antonio R. (1980). The anatomical basis of conduction aphasia. *Brain*, **103**, pp. 337-350.

Daniloff, Raymond, and Hammarberg, R.E. (1973). On defining coarticulation. *Journal of Phonetics*, **1**, pp. 239-248.

De Jong, Kenneth, Beckman, Mary E., and Edwards, Jan (1993). The interplay between prosodic structure and coarticulation. *Language and Speech*, **36**, pp. 197-212.

Dronkers, Nina F. (1996). A new brain region for coordinating speech articulation. *Nature*, **384**, pp. 159-161.

Fitts, Paul M. (1954). The information capacity of the human motor system in controlling the amplitude of movement. *Journal of Experimental Psychology*, **47**, pp. 381-391.

Fox, Peter T., Huang, Aileen, Parsons, Lawrence M., Xiong, Jin-Hu, Zamarippa, Frank, Rainey, Lacy, and Lancaster, Jack L. (2001). Location-probability profiles for the mouth region of human primary motor-sensory cortex: Model and validation. *NeuroImage*, **13**, pp. 196-209.

Fujimura, Osamu (2000). The C/D model and prosodic control of articulatory behavior. *Phonetica*, **57**, pp. 128-138.

Gay, Thomas, Ushijima, T., Hirose, H., and Cooper, Franklin S. (1974). Effects of speaking rate on labial consonant-vowel articulation. *Journal of Phonetics*, **2**, pp. 47-63.

Geschwind, Norman (1965). Disconnexion syndromes in animals and man. I. *Brain*, **88**, pp. 237-294.

Guenther, Frank H. (1992). *Neural models of adaptive sensory-motor control for flexible reaching and speaking*. Doctoral dissertation, Boston University, Boston.

Guenther, Frank H. (1994). A neural network model of speech acquisition and motor equivalent speech production. *Biological Cybernetics*, **72**, pp. 43-53.

Guenther, Frank H. (1995a). Speech sound acquisition, coarticulation, and rate effects in a neural network model of speech production. *Psychological Review*, **102**, pp. 594-621.

Guenther, Frank H. (1995b). A modeling framework for speech motor development and kinematic articulator control. *Proceedings of the XIIIth International Conference of Phonetic Sciences (*vol. 2, pp. 92-99). Stockholm, Sweden: KTH and Stockholm University.

Guenther, Frank H., Espy-Wilson, Carol Y., Boyce, Suzanne E., Matthies, Melanie L., Zandipour, Majid, and Perkell, Joseph S. (1999). Articulatory tradeoffs reduce acoustic variability during American English /r/ production. *Journal of the Acoustical Society of America*, **105**, pp. 2854-2865.

Guenther, Frank H., Hampson, Michelle, and Johnson, David (1998). A theoretical investigation of reference frames for the planning of speech movements. *Psychological Review*, **105**, pp. 611-633.

Guenther, Frank H., and Micci Barreca, Daniele (1997). Neural models for flexible control of redundant systems. In: P. Morasso and V. Sanguineti (eds.), *Self-organization, Computational Maps, and Motor Control* (pp. 383-421). Amsterdam: Elsevier-North Holland.

Hickok, Gregory, Erhard, Peter, Kassubek, Jan, Helms-Tillery, A. Kate, Naeve-Velguth, Susan, Strupp, John P., Strick, Peter L., and Ugurbil, Kamil (2000). A functional magnetic resonance imaging study of the role of left posterior superior temporal gyrus in speech production: Implications for the explanation of conduction aphasia. *Neuroscience Letters*, **287**, pp. 156-160.

Hollerbach, John M., and Suh, Ki C. (1985). Redundancy resolution of manipulators through torque optimization. *Proceedings of the IEEE International Conference on Robotics and Automation* (pp. 1016-1021). New York: IEEE.

Indefrey, Peter, and Levelt, Willem J.M. (2000). The neural correlates of language production. In M. Gazzaniga (Ed.), *The new cognitive neurosciences*, 2nd Edition (pp. 845-865). Cambridge, MA: MIT Press.

Keating, Patricia A. (1990). The window model of coarticulation: Articulatory evidence. In J. Kingston & M. E. Beckman (Eds.), *Papers in laboratory phonology I: Between the grammar and physics of speech* (pp. 451-470). Cambridge, England: Cambridge University Press.

Kent, Raymond D. (1983). The segmental organization of speech. In P. F. MacNeilage (Ed.), *The production of speech* (pp. 57-89). New York: Springer-Verlag.

Klein, Charles A., and Huang, Ching-Hsiang (1983). Review of pseudoinverse control for use with kinematically redundant manipulators. *IEEE Transactions on Systems, Man, and Cybernetics*, **SMC-13**(2), pp. 245-250.

Laboissière, Rafael, Ostry, David J., and Perrier, Pascal (1995). A model of human jaw and hyoid motion and its implications for speech production. *Proceedings of the XIIIth International Conference of Phonetic Sciences (*vol. 2, pp. 60-67). Stockholm, Sweden: KTH and Stockholm University.

Levelt, Willem J.M. (1989). *Speaking: From intention to articulation*. Cambridge, MA: MIT Press.

Levelt, Willem J.M., Praamstra, P., Meyer, Antje S., Helenius, P., and Salmelin, R. (1998). An MEG study of picture naming. *Journal of Cognitive Neuroscience*, **10**, pp. 553-567.

Levelt, Willem J.M., Roelofs, Ardi, Meyer, Antje S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences*, **22** , pp. 1-38.

Liégeois, Alain (1977). Automatic supervisory control of the configuration and behavior of multibody mechanisms. *IEEE Transactions on Systems, Man, and Cybernetics,* **SMC-7**(12), pp. 869-871.

Lindblom, Bjorn, Lubker, James, and Gay, Thomas (1979). Formant frequencies of some fixed-mandible vowels and a model of speech motor programming by predictive simulation. *Journal of Phonetics*, **7**, pp. 147-161.

Lindblom, Bjorn, and MacNeilage, Peter F. (1986). Action theory: Problems and alternative approaches. *Journal of Phonetics*, **14**, pp. 117-132.

McFarland, David H., Baum, Shari R. (1995). Incomplete compensation to articulatory perturbation. *Journal of the Acoustical Society of America*, **97**, pp. 1865-73.

Maeda, Shinji (1990). Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal tract shapes using an articulatory model. In W.J. Hardcastle and A. Marchal (Eds.), *Speech production and speech modelling* (pp. 131-149). Boston: Kluwer Academic Publishers.

Manuel, Sharon Y. (1990). The role of contrast in limiting vowel-to-vowel coarticulation in different languages. *Journal of the Acoustical Society of America*, **88**, pp. 1286-1298.

Mermelstein, Paul (1973). Articulatory model for the study of speech production. *Journal of the Acoustical Society of America*, **53**, pp. 1070-1082.

Merton, P.A. (1972). How we control the contraction of our muscles. *Scientific American*, **226**, pp. 30-37.

Miller, James D. (1989). Auditory-perceptual interpretation of the vowel. *Journal of the Acoustical Society of America*, **85**, pp. 2114-2134.

Mussa-Ivaldi, Ferdinando A., and Hogan, Neville (1991). Integrable solutions of kinematic redundancy via impedance control. *International Journal of Robotics Research*, **10**, pp. 481-491.

Ostry, David J., Gribble, Paul L., and Gracco, Vincent L. (1996). Coarticulation of jaw movements in speech production: Is context sensitivity in speech kinematics centrally planned? *Journal of Neuroscience*, **16**, pp. 1570-1579.

Paus, Tomas, Perry, David W., Zatorre, Robert J., Worsley, Keith J., and Evans, Alan C. (1996). Modulation of cerebral blood flow in the human auditory cortex during speech: Role of motor-to-sensory discharges. *European Journal of Neuroscience*, **8**, pp. 2236-2246.

Perkell, Joseph, Guenther, Frank, Lane, Harlan, Matthies, Melanie, Perrier, Pascal, Vick, Jennell, Wilhelms-Tricarico, Reiner, and Zandipour, Majid (2000). A theory of speech motor control and supporting data from speakers with normal hearing and with profound hearing loss. *Journal of Phonetics,* 28, 233-272.

Perkell, Joseph S., Matthies, Melanie L., Svirsky, Mario A., and Jordan, Michael I. (1995). Goal-based speech motor control: A theoretical framework and some preliminary data. *Journal of Phonetics,* **23**, pp. 23-35.

Perkell, Joseph S., and Nelson, W.L. (1985). Variability in production of the vowels /i/ and /a/. *Journal of the Acoustical Society of America*, **77**, pp. 1889-1895.

Picheny, Michael A., Durlach, Nathaniel I., and Braida, Louis D. (1985). Speaking clearly for the hard of hearing I: Intelligibility differences between clear and conversational speech. *Journal of Speech and Hearing Research*, **28**, pp. 96-103.

Picheny, Michael A., Durlach, Nathaniel .I., and Braida, Louis D. (1986). Speaking clearly for the hard of hearing II: Acoustic characteristics of clear and conversational speech. *Journal of Speech and Hearing Research*, **29**, pp. 434-446.

Raibert, Marc H. (1977). Motor control and learning by the state space model. Technical Report AI-M-351, Massachusetts Institute of Technology.

Rubin, Philip, Baer, Thomas, and Mermelstein, Paul (1981). An articulatory synthesizer for perceptual research. *Journal of the Acoustical Society of America*, **70**, pp. 321-328.

Saltzman, Elliot L., and Munhall, Kevin G. (1989). A dynamical approach to gestural patterning in speech production. *Ecological Psychology*, **1**, pp. 333-382.

Savariaux, Christophe, Perrier, Pascal, and Orliaguet, Jean P. (1995). Compensation strategies for the perturbation of the rounded vowel [u] using a lip tube: A study of the control space in speech production. *Journal of the Acoustical Society of America*, **98**, pp. 2428-2442.

Savariaux, Christophe, Perrier, Pascal, and Schwartz, Jean-Luc (1995). Perceptual analysis of compensatory strategies in the production of the French rounded vowel [u] perturbed by a lip tube. *Proceedings of the XIIIth International*

Congress of Phonetic Sciences (vol. 3, pp. 584-587). Stockholm, Sweden: KTH and Stockholm University.

Sereno, Joan A., and Lieberman, Philip (1987). Developmental aspects of lingual coarticulation. *Journal of Phonetics*, **15**, pp. 247-257.

Stein, Paul S.G., Mortin, L.I., and Robertson, G.A. (1986). The forms of a task and their blends. In S. Grillner, P.S.G. Stein, D.G. Stuart, H. Forssberg, and R.M. Herman (eds.), *Neurobiology of Vertebrate Locomotion* (pp. 201-216). London: Macmillan.

Stone, Maureen (1991). Toward a model of three-dimensional tongue movement. *Journal of Phonetics*, **19**, pp. 309-320.

Thompson, A.E., and Hixon, T.J. (1979). Nasal air flow during normal speech production. *Cleft Palate Journal*, **16**, pp. 412-420.

Whalen, Douglas H. (1990). Coarticulation is largely planned. *Journal of Phonetics*, **18**, pp. 3-35.

Whitney, D.E. (1969). Resolved motion rate control of manipulators and human prostheses. *IEEE Transactions on Man-Machine Systems*, **MMS-10**(2), pp. 47-53.

Wilhelms-Tricarico, Reiner (1995). Physiological modeling of speech production: Methods for modeling soft-tissue articulators. *Journal of the Acoustical Society of America*, **97**, pp. 3085-3098.

Wilhelms-Tricarico, Reiner (1996). A biomechanical and physiologically-based vocal tract model and its control. *Journal of Phonetics*, **24**, pp. 23-38.

Wise, Richard J., Greene, J., Büchel, Christian, and Scott, Sophie K. (1999). Brain regions involved in articulation. *Lancet*, **353**, pp. 1057-1061.

Woodworth, Robert S. (1899). The accuracy of voluntary movement. *Psychological Review*, **3**, pp. 1-114.

Wright, Charles E. (1990). Generalized motor programs: Reexamining claims of effector independence in writing. In Jeannerod, M. (ed.), *Attention and performance XIII: Motor representation and control* (pp. 294-320). Hillsdale, NJ: Erlbaum.

Acknowledgements

This research was supported by the National Institute on Deafness and other Communication Disorders (NIH grants R01 DC02852, F. Guenther PI; R01 DC01925, R01 DC03007, J. Perkell, PI).