Published in final edited form as:
Brain Lang. 2006 March ; 96(3): 280–301.

Neural Modeling and Imaging of the Cortical Interactions Underlying Syllable Production

Frank H. Guenther[1,2,3], **Satrajit S. Ghosh**[1], and **Jason A. Tourville**[1]

1 *Department of Cognitive and Neural Systems, Boston University, 677 Beacon Street, Boston, MA, 02215, Telephone: (617) 353-5765, Fax Number: (617) 353-7755, Email: guenther@cns.bu.edu*

2 *Speech and Hearing Bioscience and Technology Program Harvard University/Massachusetts Institute of Technology Cambridge, MA 02139*

3 *Athinoula A. Martinos Center for Biomedical Imaging Massachusetts General Hospital Charlestown, MA 02129*

Abstract

This paper describes a neural model of speech acquisition and production that accounts for a wide range of acoustic, kinematic, and neuroimaging data concerning the control of speech movements. The model is a neural network whose components correspond to regions of the cerebral cortex and cerebellum, including premotor, motor, auditory, and somatosensory cortical areas. Computer simulations of the model verify its ability to account for compensation to lip and jaw perturbations during speech. Specific anatomical locations of the model's components are estimated, and these estimates are used to simulate fMRI experiments of simple syllable production.

Keywords

speech production; model; fMRI; Broca's area; premotor cortex; motor cortex; speech acquisition; sensorimotor learning; neural transmission delays

Introduction

The advent of functional brain imaging techniques that are safe for use on human subjects has led to an explosion in the amount of data concerning brain activity during speech and language tasks. The current article details a neural model of speech production that provides a conceptual and computational framework for interpreting many of these datasets. The model is a neural network model of speech acquisition and production, called the DIVA model (Directions Into Velocities of Articulators), that utilizes a babbling cycle to learn to control movements of simulated speech articulators in order to produce phoneme strings. Over the past decade, our laboratory has used numerical simulations to show how the model provides a relatively simple, unified account of a very wide range of speech production phenomena, including motor equivalence, contextual variability, anticipatory and carryover coarticulation, velocity/distance relationships, speaking rate effects, and speaking skill acquisition (e.g., Guenther, 1994; Guenther, 1995; Guenther, Hampson, & Johnson, 1998; Guenther & Ghosh, 2003; Nieto-Castanon, in press). Predictions concerning speech production in normal adults have been drawn from the model and tested using electromagnetic articulometry (e.g., Guenther et al., 1999; Perkell et al., 2004). The model has been used to account for issues in child development

Supported by the National Institute on Deafness and other Communication Disorders (R01 DC02852, F. Guenther PI; S. Ghosh also supported by R01 DC01925, J. Perkell PI).

(e.g., Guenther, 1995), including a demonstration of its ability to deal with the dramatic changes in size and shape of the speech articulators that take place during the first three years of life (Callan, Kent, Guenther, & Vorperian, 2000). The model has also been used to investigate the role of auditory feedback in speech production in normally hearing individuals, deaf individuals, and individuals who have recently regained some hearing through the use of cochlear implants (Perkell et al., 2000), and to investigate stuttering (Max, Guenther, Gracco, Ghosh, & Wallace, 2004). Because the DIVA model is defined as a neural network, its components can be interpreted in terms of brain function in a straightforward way. The model thus provides an ideal framework for interpreting data from functional imaging studies of the human brain during speech tasks. Preliminary associations of the model's components with specific brain regions have been presented elsewhere (e.g., Guenther, 1998; Guenther, 2001; Guenther et al., 2003); a primary goal of the current paper is to provide a more thorough treatment of the hypothesized neural bases of the model's components.

A second purpose of the current work is to extend the model to incorporate realistic neural processing delays, and therefore more realistically address the issue of combining feedforward and feedback control strategies. Earlier versions of the DIVA model effectively assumed instantaneous transmission of neural signals. However the nervous system must cope with potentially destabilizing delays in the control of articulator movements. For example, a motor command generated in the primary motor cortex will typically take 40 ms or more before it effects movement of the associated speech articulator. Similarly, sensory information from the articulators and cochlea are delayed by tens of ms before they reach the primary sensory cortices. These transmission delays can be very problematic for a system that must control the rapid articulator movements underlying speech. Most adults can pronounce the word "dilapidated" in less than one second; this word requires 10 transitions between phonemes, with each transition taking less than 100ms to complete. A purely feedback-based control system faced with the delays mentioned above would not be able to stably produce speech at this rate. Instead, our speech production system must supplement feedback control with feedforward control mechanisms. In this article we address the integration of feedback and feedforward control subsystems in the control of speech movements with realistic processing delays, and we provide model simulations of perturbation studies that probe the temporal response properties of feedback control mechanisms.

Several aspects of the DIVA model differentiate it from other models in the speech production literature (e.g., Levelt, Roelofs, & Meyer, 1999; Saltzman & Munhall, 1989; Morasso, Sanguineti, & Frisone, 2001; Westermann & Reck, 2004). Whereas the Levelt, Roelofs, & Meyer (1999) model focuses on linguistic and phonological computations down to the syllable level, the DIVA model focuses on the sensorimotor transformations underlying the control of articulator movements. Thus, the DIVA model focuses on speech control at the syllable and lower motor levels. The task dynamic model of Saltzman et al. (1989) is a computational model that provides an alternative account of the control of articulator movements. However, unlike the DIVA model its components are not associated with particular brain regions, neuron types, or synaptic pathways. Of current biologically plausible neural network models of speech production (e.g., Morasso et al., 2001; Westermann et al., 2004), the DIVA model is the most thoroughly defined and tested, and it is unique in using a pseudoinverse-style control scheme (from which the model's name is derived) that has been shown to provide accurate accounts of human articulator kinematic data (e.g., Guenther et al., 1998; Guenther et al., 1999; Nieto Castanon et al., in press). It is also unique in using a combination of feedback and feedforward control mechanisms (as described in the current article), as well as embodying a *convex region theory* for the targets of speech that has been shown to provide a unified account of a wide body of speech acoustic, kinematic, and EMG data (Guenther, 1995).

An overview of the DIVA model and description of its components are provided in the next section. Subsequent sections relate the model's components to regions of the cerebral cortex and cerebellum, including mathematical characterizations of the model's components and treatment of the relevant neurophysiological literature. Computer simulations of the model producing normal and perturbed speech are then presented, followed by a more precise account of fMRI activations measured during simple syllable production in terms of the model's cell activities.

OVERVIEW OF THE DIVA MODEL

The DIVA model, schematized in Figure 1, is an adaptive neural network that learns to control movements of a simulated vocal tract, or articulatory synthesizer (a modified version of the synthesizer described by Maeda, 1990), in order to produce words, syllables, or phonemes. The neural network takes as input a speech sound string and generates as output a time sequence of articulator positions that command movements of the simulated vocal tract. Each block in the model schematic (Figure 1) corresponds to a set of neurons that constitute a *neural representation*. In this article, the term *map* will be used to refer to such a set of cells. The term *mapping* will be used to refer to a transformation from one neural representation to another (arrows in Figure 1), assumed to be carried out by filtering cell activations in one map through synapses projecting to another map. The synaptic weights are tuned during a babbling phase in which random movements of the speech articulators provide tactile, proprioceptive, and auditory feedback signals that are used to learn the mappings between different neural representations. After babbling, the model can quickly learn to produce new sounds from audio samples provided to it, and it can produce arbitrary combinations of the sounds it has learned.

In the model, production of a phoneme or syllable starts with activation of a speech sound map cell, hypothesized to lie in ventral premotor cortex, corresponding to the sound to be produced. After a speech sound map cell has been activated, signals from premotor cortex travel to the auditory and somatosensory cortical areas through tuned synapses that encode sensory expectations for the sound. Additional synaptic projections from speech sound map cells to the model's motor cortex (both directly and via the cerebellum) form a feedforward motor command.

The synapses projecting from the premotor cortex to auditory cortical areas encode an expected auditory trace for each speech sound. They can be tuned while listening to phonemes and syllables from the native language or listening to correct self-productions. After learning, these synapses encode a spatiotemporal target region for the sound in auditory coordinates. During production of the sound, this target region is compared to the current auditory state, and any discrepancy between the target and the current state, or auditory error, will lead to a command signal to motor cortex that acts to correct the discrepancy via projections from auditory to motor cortical areas.

Synapses projecting from the premotor cortex to somatosensory cortical areas encode the expected somatic sensation corresponding to the active syllable. This spatiotemporal somatosensory target region is estimated by monitoring the somatosensory consequences of producing the syllable over many successful production attempts. Somatosensory error signals are then mapped to corrective motor commands via pathways projecting from somatosensory to motor cortical areas.

Feedforward and feedback control signals are combined in the model's motor cortex. Feedback control signals project from sensory error cells to the motor cortex as described above. These projections are tuned during babbling by monitoring the relationship between sensory signals and the motor commands that generated them. The feedforward motor command is hypothesized to project from ventrolateral premotor cortex to primary motor cortex, both

directly and via the cerebellum. This command can be learned over time by averaging the motor commands from previous attempts to produce the sound.

The following sections present the model's components in further detail, including a mathematical characterization of the cell activities in the cortical maps and a treatment of relevant neuroanatomical and neurophysiological findings (with a more detailed neurophysiological treatment provided in Appendix A). For purposes of exposition, the model's premotor and motor cortical representations will be treated first, followed by treatments of the feedback and feedforward control subsystems.

MOTOR AND PREMOTOR REPRESENTATIONS

Premotor Cortex Speech Sound Map

Each cell in the model's ***speech sound map***, hypothesized to correspond to neurons in the left ventral premotor cortex and/or posterior Broca's area[1], represents a different speech sound[2]. A "speech sound" is defined here as a phoneme, syllable, word, or short phrase that is frequently encountered in the native language and therefore has associated with it a stored motor program for its production. For example, we expect all phonemes and frequent syllables of a language to be represented by unique speech sound map cells. In contrast, we expect that infrequent syllables do not have stored motor programs associated with them; instead we expect they are produced by sequentially instating the motor programs of the phonemes (or other sub-syllabic sound chunks, such as demisyllables cf. Fujimura & Lovins, 1978) that form the syllable. In terms of our model, infrequent syllables are produced by sequentially activating the speech sound map cells corresponding to the smaller sounds that make up the syllable.

Speech sound map cells are hypothesized to lie in ventral premotor cortex because of their functional correspondence with "mirror neurons." Mirror neurons are so termed because they respond both during an action and while viewing (or hearing) that action performed by another animal or person (Rizzolatti et al., 1996; Kohler et al., 2002). These cells have been shown to code for complex actions such as grasping rather than the individual movements that comprise an action (Rizzolatti et al, 1988). Neurons within the speech sound map are hypothesized to embody similar properties: activation during speech production drives complex articulator movement, and activation during speech perception tunes connections between the speech sound map and sensory cortex (described further below; see Arbib, in press for a different view of the role of mirror neurons in language.)

Demonstrations of mirror neurons in humans have implicated left precentral gyrus for grasping actions (Tai, Scherfler, Brooks, Sawamoto, & Castiello, 2004), and left hemisphere opercular inferior frontal gyrus for finger movements (Iacoboni et al., 1999). Recently, mirror neurons related to communicative mouth movements have been found in monkey area F5 (Ferrari et al., 2003) immediately lateral to their location for grasping movements (Di Pellegrino, Fadiga, Fogassi, Gallese, & Rizzolatti, 1992). This area has been proposed to correspond to the caudal portion of ventral inferior frontal gyrus (Brodmann's area 44) in the human (see Rizzolatti & Arbib, 1998). We therefore propose that the speech sound map cells lie in ventral lateral premotor areas of the left hemisphere[3], including posterior portions of the inferior frontal gyrus.

[1]In this paper, we use the term Broca's area to refer to the inferior frontal gyrus pars opercularis (posterior Broca's area) and pars triangularis (anterior Broca's area). Due to the large amount of inter-subject variability in the location of the ventral precentral sulcus as measured in stereotactic coordinates, it is difficult to differentiate the ventral premotor cortex and posterior Broca's area in fMRI or PET studies that involve averaging across subjects using standard normalization techniques (see Nieto-Castanon, Ghosh, Tourville, & Guenther, 2003 and Tomaiuolo et al., 1999 for related discussions).

[2]Although each sound is represented by a single speech sound map cell in the model, it is expected that premotor cortex sound maps in the brain involve distributed representations of each speech sound. These distributed representations would be more robust to potential problems such as cell death and would allow greater generalizability of learned motor programs to new sounds. However these topics are beyond the scope of the current article.

The equation governing speech sound map cell activation in the model is:

$$P_i(t) = 1 \;\; if \; i^{th} \; sound \; is \; being \; produced \; or \; perceived$$
$$P_i(t) = 0 \;\; otherwise \qquad (1)$$

Each time a new speech sound is presented to the model (as an acoustic sample) for learning, a new cell is recruited into the speech sound map to represent that sound. There are several aspects to this learning, described further below. After the sound has been learned, activation of the speech sound map cell leads to production of the corresponding sound via the model's feedforward and feedback subsystems.

The model's speech sound map cells can be interpreted as forming a "mental syllabary" as described by Levelt and colleagues (e.g., Levelt & Wheeldon, 1994; Levelt et al., 1999). Levelt et al. (1999) describe the syllabary as a "repository of gestural scores for the frequently used syllables of the language" (p. 5). According to our account, higher-level brain regions involved in phonological encoding of an intended utterance (e.g., anterior Broca's area) sequentially activate speech sound map cells that correspond to the syllables to be produced. The activation of these cells leads to the readout of feedforward motor commands to the primary motor cortex (see *Feedforward Control Subsystem* below), as well as a feedback control command if there is any error during production (see *Feedback Control Subsystem*). The feedforward command emanating from a speech sound map cell can be thought of as the "motor program" or "gestural score", i.e., a time sequence of motor gestures used to produce the corresponding speech sound (cf. Browman & Goldstein, 1989).

According to the model, when an infant listens to a speaker producing a new speech sound, a previously unused speech sound map cell becomes active, and projections from this cell to auditory cortical areas are tuned to represent the auditory signal corresponding to that sound. The projections from the premotor speech sound map cells to the auditory cortex represent a target auditory trace for that sound; this auditory target is subsequently used in the production of the sound (see *Feedback Control Subsystem* below for details), along with feedforward commands projecting from the speech sound map cell to the motor cortex (detailed in *Feedforward Control Subsystem* below).

Motor Cortex Velocity and Position Maps

According to the model, feedforward and feedback-based control signals are combined in motor cortex. Three distinct subpopulations (maps) of motor cortical cells are thought to be involved in this process: one population representing positional commands to the speech articulators, one representing velocity commands originating from the feedforward control subsystem, and one representing velocity commands originating from the feedback control subsystem.

Cells in the model's ***motor cortex position map*** correspond to "tonic" cells found in motor cortex electrophysiological studies in monkeys (e.g., Kalaska, Cohen, Hyde, & Prud'homme, 1989). Their activities at time *t* are represented by the vector *M(t)*. The motor position cells are formed into antagonistic pairs, with each pair representing a position command for one of the eight model articulators. Thus M(t) is a 16-dimensional vector, and it is governed by the following equation:

[3]All cell types in the model other than the speech sound map cells are thought to exist bilaterally in the cerebral cortex.

$$M(t) = M(0) + \alpha_{ff}\int_0^t M_{Feedforward}(t)g(t)dt + \alpha_{fb}\int_0^t M_{Feedback}(t)g(t)dt \quad (2)$$

where $M(0)$ is the initial configuration of the vocal tract when starting an utterance, α_{fb} and α_{ff} are parameters that determine how much the model is weighted toward feedback control and feedforward control[4], respectively, and $g(t)$ is a speaking rate signal that is 0 when not speaking and 1 when speaking at a maximum rate. The 16-dimensional vectors $M_{Feedforward}(t)$ and $M_{Feedback}(t)$ constitute the model's ***motor cortex velocity maps*** and correspond to "phasic" cells found in electrophysiological studies in monkeys (e.g., Kalaska et al., 1989). $M_{Feedforward}(t)$ encodes a feedforward control signal projecting from premotor cortex and the cerebellum, and $M_{Feedback}(t)$ encodes a feedback control signal projecting from sensory cortical areas; the sources of these command signals are discussed further in later sections (*Feedback Control Subsystem* and *Feedforward Control Subsystem*).

The model's motor position map cells produce movements in the model's articulators according to the following equation:

$$Artic(t) = f_{MAr}(M(t - \tau_{MAr})) + Pert(t) \quad (3)$$

where f_{MAr} is a simple function relating the motor cortex position command to the Maeda parameter values (transforming each antagonistic pair into a single articulator position value), τ_{MAr} is the time it takes for a motor command to have its effect on the articulatory mechanism, and *Pert* is the effect of external perturbations on the articulators if such perturbations are applied (see *Computer Simulations of the Model* below). The eight-dimensional vector *Artic* does not correspond to any cell activities in the model; it corresponds instead to the physical positions of the eight articulators[5] in the Maeda articulatory synthesizer (Maeda, 1990). The resulting vocal tract area function is converted into a digital filter that is used to synthesize an acoustic signal that forms the output of the model (e.g., Maeda, 1990).

Roughly speaking, the delay τ_{MAr} in Equation 3 corresponds to the time it takes for an action potential in a motor cortical cell to affect the length of a muscle via a subcortical motoneuron. This time can be broken into two components: (1) the delay between motor cortex activation and activation of a muscle as measured by EMG, and (2) the delay between EMG onset and muscle length change. Regarding the former, Meyer, Werhahn, Rothwell, Roericht, and Fauth (1994) measured the latency of EMG responses to transcranial magnetic stimulation of the face area of motor cortex in humans and found latencies of 11–12 ms for both ipsilateral and contralateral facial muscles. Regarding the latter, time delays between EMG onset and onset of the corresponding articulator acceleration of approximately 30 ms have been measured in the posterior genioglossus muscle of the tongue (Majid Zandipour and Joseph Perkell, personal communication); this estimate is in line with a more thorough investigation of bullfrog muscles which showed average EMG to movement onset latencies of approximately 24 ms in hip extensor muscles, with longer latencies occurring in other leg muscles (Olson & Marsh, 1998). In keeping with these results, we use τ_{MAr} = 42 ms in the simulations reported below.

[4]Under normal circumstances, both α_{fb} and α_{ff} are assumed to be 1. However, certain motor disorders may be associated with an inappropriate balance between feedforward and feedback control. For example, stuttering can be induced in the model by using an inappropriately low value of α_{ff} (see Guenther & Ghosh, 2003).

[5]The eight articulators in the modified version of the Maeda synthesizer used herein correspond approximately to jaw height, tongue shape, tongue body position, tongue tip position, lip protrusion, larynx height, upper lip height and lower lip height,. These articulators were based on a modified principal components analysis of midsagittal vocal tract outlines, and each articulator can be varied from −3.5 to +3.5 standard deviations from a neutral configuration.

When an estimate of EMG onset latency is needed in the simulations, we use a 12 ms estimate from motor cortical cell activation to EMG onset based on Meyer et al. (1994).

The next two sections describe the feedback and feedforward control subsystems that are responsible for generating the motor commands $M_{Feedback}(t)$ and $M_{Feedforward}(t)$.

FEEDBACK CONTROL SUBSYSTEM

The feedback control subsystem in the DIVA model (blue portion of Figure 1) carries out the following functions when producing a learned sound. First, activation of the speech sound map cell corresponding to the sound in the model's premotor cortex leads to readout of learned auditory and somatosensory targets for that sound. These targets take the form of temporally varying regions in the auditory and somatosensory spaces, as described below. The current auditory and somatosensory states, available through sensory feedback, are compared to these targets in the higher-order auditory and somatosensory cortices. If the current sensory state falls outside of the target region, an error signal arises in the higher-order sensory cortex. These error signals are then mapped into appropriate corrective motor commands via learned projections from the sensory error cells to the motor cortex.

The following paragraphs detail these processes, starting with descriptions of the auditory and somatosensory state maps, continuing with a treatment of the auditory and somatosensory targets for a speech sound, and concluding with a description of the circuitry involved in transforming auditory and somatosensory error signals into corrective motor commands.

Auditory State Map

In the model, the acoustic state is determined from the articulatory state as follows:

$$Acoust(t) = f_{ArAc}(Artic(t)) \quad (4)$$

where f_{ArAc} is the transformation performed by Maeda's articulatory synthesis software. The vector *Acoust(t)* does not correspond to brain cell activities; instead it corresponds to the physical acoustic signal resulting from the current articulator configuration.

The model includes an **a*uditory state map*** that corresponds to the representation of speech-like sounds in auditory cortical areas (BA 41, 42, 22). The activity of these cells is represented as follows:

$$Au(t) = f_{AcAu}(Acoust(t - \tau_{AcAu})) \quad (5)$$

where *Au(t)* is a vector of auditory state map cell activities, f_{AcAu} is a function that transforms an acoustic signal into the corresponding auditory cortical map representation, and τ_{AcAu} is the time it takes an acoustic signal transduced by the cochlea to make its way to the auditory cortical areas. Regarding τ_{AcAu}, Schroeder and Foxe (2002) measured the latency between onset of an auditory stimulus and responses in higher-order auditory cortical areas posterior to A1 and a superior temporal polysensory (STP) area in the dorsal bank of the superior temporal sulcus. They noted a response latency of approximately 10 ms in the posterior auditory cortex and 25 ms in STP. Based in part on these numbers, we use an estimate of τ_{AcAu} = 20 ms in the simulations reported below.

Regarding f_{AcAu}, we have used a variety of different auditory representations in the model, including formant frequencies, log formant ratios (e.g., Miller, 1989), and wavelet-based transformations of the acoustic signal. Simulations using these different auditory spaces have yielded similar results in most cases. In the computer simulations reported below, we use a

formant frequency representation in which *Au(t)* is a three-dimensional vector whose components correspond to the first three formant frequencies of the acoustic signal.

Somatosensory State Map

The model also includes a ***somatosensory state map*** that corresponds to the representation of speech articulators in somatosensory cortical areas (BA 1,2,3,40,43):

$$S(t) = f_{ArS}(Artic(t - \tau_{ArS})) \quad (6)$$

where *S(t)* is a 22-dimensional vector of somatosensory state map cell activities, f_{ArS} is a function that transforms the current state of the articulators into the corresponding somatosensory cortical map representation, and τ_{ArS} is the time it takes somatosensory feedback from the periphery to reach higher-order somatosensory cortical areas. Regarding τ_{ArS}, O'Brien, Pimpaneau, and Albe-Fessard (1971) measured evoked potentials in somatosensory cortex induced by stimulation of facial nerves innervating the lips, jaw, tongue, and larynx in anesthetized monkeys. They report typical latencies of approximately 5–20 ms, though some somatosensory cortical cells had significantly longer latencies, on the order of 50 ms. Schroeder and Foxe (2002) noted latencies of approximately 10 ms in inferior parietal sulcus to somatosensory stimulation (electrical stimulation of a hand nerve). Based on these results, we use an estimate of τ_{ArS} = 15 ms in the simulations reported below.

The function f_{ArS} transforms the articulatory state into a 22-dimensional somatosensory map representation *S(t)* as follows. The first 16 dimensions of *S(t)* correspond to proprioceptive feedback representing the current positions of the 8 Maeda articulators, each represented by an antagonistic pair of cells as in the motor representation. In other words, the portion of f_{ArS} that determines the first 16 dimensions of *S(t)* is basically the inverse of $f_{Mar.}$ The remaining 6 dimensions correspond to tactile feedback, consisting of palatal and labial tactile information derived from the first five Maeda articulatory parameters using a simple modification of the mapping described by Schwartz and Boë (Schwartz & Boë, 2000).

Motor-to-sensory pathways encode speech sound targets

We hypothesize that axonal projections from speech sound map cells in the frontal motor cortical areas (lateral BA 6 and 44) to higher-order auditory cortical areas[6] in the superior temporal gyrus (BA 22) carry auditory targets for the speech sound currently being produced. That is, these projections predict the sound of the speaker's own voice while producing the sound based on auditory examples from other speakers producing the sound, as well as one's own previous correct productions. Furthermore, projections from the speech sound map cells to higher-order somatosensory cortical areas in the anterior supramarginal gyrus and surrounding cortex (BA 40; perhaps also portions of BA 1, 2, 3, and 43) are hypothesized to carry target (expected) tactile and proprioceptive sensations associated with the sound currently being produced. These expectations are based on prior successful attempts to produce the sound, though we envision the possibility that some aspects of the somatosensory targets might be learned by infants when they view a speaker (e.g., by storing the movement of the lips for a bilabial).

[6]Although currently treated as a single set of synaptic weights in the model, it is possible that this mapping may include a trans-cerebellar contribution (motor cortex -> pons -> cerebellum -> thalamus -> higher-order auditory cortex) in addition to a cortico-cortical contribution. We feel that current data do not definitively resolve this issue. The weight matrix z_{PAu} (as well as z_{PS}, z_{SM}, and z_{AuM}, defined below) can thus be considered as (possibly) combining cortico-cortical and trans-cerebellar synaptic projections. We consider the evidence for a trans-cerebellar contribution to the weight matrix z_{PM}, which encodes a feedforward command between the premotor and motor cortices as described in the next section, to be much stronger.

The auditory and somatosensory targets take the form of multidimensional regions, rather than points, that can vary with time, as schematized in Figure 2. The use of target regions is an important aspect of the DIVA model that provides a unified explanation for a wide range of speech production phenomena, including motor equivalence, contextual variability, anticipatory coarticulation, carryover coarticulation, and speaking rate effects (see Guenther, 1995 for details).

In the computer simulations, the auditory and somatosensory targets for a speech sound are encoded by the weights of the synapses projecting from the premotor cortex (specifically, from the speech sound map cell representing the sound) to cells in the higher-order auditory and somatosensory cortices, respectively. The synaptic weights encoding the auditory target for a speech sound are denoted by the matrix $z_{PAu}(t)$, and the weights encoding the somatosensory target are denoted by the matrix $z_{PS}(t)$. These weight matrices are "spatiotemporal" in that they encode target regions for each point in time from the start of production to the end of production of the speech sound they encode. That is, each column of the weight matrix represents the target at one point in time, and there is a different column for every 1 ms of the duration of the speech sound.

It is hypothesized that the weights $z_{PAu}(t)$become tuned when an infant listens to examples of a speech sound, e.g. as produced by his/her parents. In the current model the weights are algorithmically tuned[7] by presenting the model with an audio file containing a speech sound produced by an adult male. The weights $z_{PAu}(t)$ encoding that sound are then adjusted so that they encode upper and lower bounds for each of the first three formant frequencies at 1 ms intervals for the duration of the utterance.

It is further hypothesized that the weights $z_{PS}(t)$ become tuned during correct self-productions of the corresponding speech sound. Note that this occurs after learning of the auditory target for the sound since the auditory target can be learned simply by monitoring a sound spoken by someone else Many aspects of the somatosensory target, however, require monitoring of correct self-productions of the sound, which are expected to occur after (and possibly during) the learning of feedforward commands for producing the sound (described in the next section). In the model the weights $z_{PS}(t)$are adjusted to encode upper and lower bounds for each somatosensory dimension at 1 ms intervals for the duration of the utterance.

In the motor control literature, it is common to refer to internal estimates of the sensory consequences of movements as "forward models". The weight matrices $z_{PAu}(t)$ and $z_{PS}(t)$ are examples of forward models in this sense. Although not currently implemented in the model, we also envision the possibility that lower-level forward models are implemented via projections from the primary motor cortex to the primary somatosensory and auditory cortices, in parallel with the $z_{PAu}(t)$ and $z_{PS}(t)$projections from premotor cortex to higher-order somatosensory and auditory cortices. Such projections would not be expected to significantly change the model's functional properties.

Auditory and somatosensory error maps

The sensory target regions for the current sound are compared to incoming sensory information in the model's higher-order sensory cortices. If the current sensory state is outside the target region error signals arise, and these error signals are mapped into corrective motor commands.

[7]By "algorithmically" we mean that a computer algorithm performs the computation without a corresponding mathematical equation, unlike other computations in the model which use numerical integration of the specified differential equations. This approach is taken to simplify the computer simulations; biologically plausible alternatives have been detailed elsewhere (e.g., Guenther, 1994; Guenther, 1995; Guenther et al., 1998). See Appendix B for further details.

The model's ***auditory error map*** encodes the difference between the auditory target region for the sound being produced and the current auditory state as represented by $Au(t)$. The activity of the auditory error map cells (ΔAu) is defined by the following equation:

$$\Delta Au(t) = Au(t) - P(t - \tau_{PAu})z_{PAu}(t) \quad (7)$$

where τ_{PAu} is the propagation delay for the signals from premotor cortex to auditory cortex (assumed to be 3ms in the simulations[8]), and $z_{PAu}(t)$ are synaptic weights that encode auditory expectations for the sound being produced. The auditory error cells become active during production if the speaker's auditory feedback of his/her own speech deviates from the auditory target region for the sound being produced.

The projections from premotor cortex represented in Equation 6 cause inhibition[9] of auditory error map cells. Evidence for inhibition of auditory cortical areas in the superior temporal gyrus during one's own speech comes from several different sources, including recorded neural responses during open brain surgery (Creutzfeldt, Ojemann, & Lettich, 1989a; Creutzfeldt, Ojemann, & Lettich, 1989b), MEG measurements (Numminen et al., 1999a; Numminen et al., 1999b), and PET measurements (Wise et al., 1999). Houde, Nagarajan, Sekihara, and Merzenich (2002) note that auditory evoked responses measured with MEG were smaller to self-produced speech than when the same speech was presented while the subject was not speaking, while response to a gated noise stimulus was the same in the presence or absence of self-produced speech. The authors concluded that "during speech production, the auditory cortex (1) attenuates its sensitivity and (2) modulates its activity as a function of the expected acoustic feedback" (p. 1125), consistent with the model.

The model's ***somatosensory error map*** codes the difference between the somatosensory target region for a speech sound and the current somatosensory state:

$$\Delta S(t) = S(t) - P(t - \tau_{PS})z_{PS}(t) \quad (8)$$

where τ_{PS} is the propagation delay from premotor cortex to somatosensory cortex (3 ms in the simulations), and the weights $z_{PS}(t)$ encode somatosensory expectations for the sound being produced. The somatosensory error cells become active during production if the speaker's somatosensory feedback from the vocal tract deviates from the somatosensory target region for the sound being produced. To our knowledge, no studies have looked for an inhibitory effect in the supramarginal gyrus during speech production, although this brain region has been implicated in phonological processing for speech perception (e.g., Caplan, Gow, & Makris, 1995; Celsis et al., 1999), and speech production (Geschwind, 1965; Damasio & Damasio, 1980).

Converting sensory errors into corrective motor actions

In the model, production errors represented by activations in the auditory and/or somatosensory error maps get mapped into corrective motor commands through learned pathways projecting from the sensory cortical areas to the motor cortex. These projections form a feedback control signal that is governed by the following equation:

$$M_{Feedback}(t) = \Delta Au(t - \tau_{AuM})z_{AuM} + \Delta S(t - \tau_{SM})z_{SM} \quad (9)$$

[8]Long-range cortico-cortical signal transmission delays are assumed to be 3 ms in the simulations, a rough estimate based on the assumption of 1–2 chemical synapses between cortical areas.

[9]These inhibitory connections are thought to involve excitatory projections from pyramidal cells in the lateral premotor cortex to local inhibitory interneurons in the auditory and somatosensory cortices.

where z_{AuM} and z_{SM} are synaptic weights that transform directional sensory error signals into motor velocities that correct for these errors, and τ_{AuM} and τ_{SM} are cortico-cortical transmission delays (3 ms in the simulations). The model's name, DIVA, derives from this mapping from sensory ***d***irections ***i***nto ***v***elocities of ***a***rticulators. Mathematically speaking, the weights z_{AuM} and z_{SM} approximate a pseudoinverse of the Jacobian of the function relating articulator positions (M) to the corresponding sensory state (Au, S; see Guenther et al., 1998 for details). Though calculated algorithmically in the current implementation (see Appendix B for details), these weights are believed to be tuned during an early babbling stage by monitoring the relationship between movement commands and their sensory consequences (see Guenther, 1995 and Guenther, 1998 for simulations involving learning of the weights). These synaptic weights effectively implement what is sometimes referred to as an "inverse model" in the motor control literature since they represent an inverse kinematic transformation between desired sensory consequences and appropriate motor actions.

The model implicitly predicts that auditory or somatosensory errors will be corrected via the feedback-based control mechanism, and that these corrections will eventually become coded into the feedforward controller if the errors are consistently encountered (see next section for learning in the feedforward control subsystem). This would be the case if a systematic auditory perturbation was applied (e.g, a shifting of one or more of the formant frequencies in real time) or a consistent somatosensory perturbation is applied (e.g., a perturbation to the jaw). Relatedly, Houde and Jordan (1998) modified the auditory feedback of speakers (specifically, shifting the first two formant frequencies of the spoken utterances and feeding this shifted auditory information back to the speaker with a time lag of approximately 16 ms) and noted that the speakers compensated for the shifted auditory feedback over time. Tremblay, Shiller, and Ostry (2003) performed an experiment in which jaw motion during syllable production was modified by application of a force to the jaw which did not measurably affect the acoustics of the syllable productions. Despite no change in the acoustics, subjects compensated for the jaw force, suggesting that they were using somatosensory targets such as those represented by $z_{PS}(t)$ in the DIVA model. The DIVA model provides a mechanistic account of these sensorimotor adaptation results.

FEEDFORWARD CONTROL SUBSYSTEM

According to the model, projections from premotor to primary motor cortex, supplemented by cerebellar projections (see Figure 1), constitute feedforward motor commands. The primary motor and premotor cortices are well-known to be strongly interconnected (e.g., Passingham, 1993;Krakauer & Ghez, 1999). Furthermore, the cerebellum is known to receive input via the pontine nuclei from premotor cortical areas, as well as higher-order auditory and somatosensory areas that can provide state information important for choosing motor commands (e.g., Schmahmann & Pandya, 1997), and projects heavily to the motor cortex (e.g., Middleton & Strick, 1997). We believe these projections are involved in the learning and maintenance of feedforward commands for the production of syllables.

Before the model has any practice producing a speech sound, the contribution of the feedforward control signal to the overall motor command will be small since it will not yet be tuned. Therefore, during the first few productions, the primary mode of control will be feedback control. During these early productions, the feedforward control system is "tuning itself up" by monitoring the motor commands generated by the feedback control system (see also Kawato & Gomi, 1992). The feedforward system gets better and better over time, all but eliminating the need for feedback-based control except when external constraints are applied to the articulators (e.g., a bite block) or auditory feedback is artificially perturbed. As the speech articulators get larger with growth, the feedback-based control system provides corrective commands that are eventually subsumed into the feedforward controller. This allows the

feedforward controller to stay properly tuned despite dramatic changes in the sizes and shapes of the speech articulators over the course of a lifetime.

The feedforward motor command for production of a sound is represented in the model by the following equation:

$$M_{Feedforward}(t) = P(t)z_{PM}(t) - M(t) \quad (10)$$

The weights $z_{PM}(t)$ encode the feedforward motor command for the speech sound being produced (assumed to include both cortico-cortical and trans-cerebellar contributions). This command is learned over time by incorporating the corrective motor commands from the feedback control subsystem on the previous attempt into the new feedforward command (see Appendix B for details).

As mentioned above, once an appropriate feedforward command sequence has been learned for a speech sound, this sequence will successfully produce the sound with very little, if any, contribution from the feedback subsystem, which will automatically become disengaged since no sensory errors will arise during production unless unexpected constraints are placed on the articulators or the auditory signal is perturbed.

COMPUTER SIMULATIONS OF THE MODEL

This section describes new computer simulations that illustrate the model's ability to learn to produce new speech sounds in the presence of neural and biomechanical processing delays, as well as to simulate the patterns of lip, jaw, and tongue movements seen in articulator perturbation experiments. Introducing perturbations during a speech task and observing the system response provides information about the nature of the controller. In particular, the time course and movement characteristics of the response can provide a window into the control processes, including neural transmission delays and the nature of the transformation between sensory and motor representations.

The simulations utilize Equations 1–10, with the delay parameters in the equations set to the values indicated below each equation. Prior to the simulations described below, the model's synaptic weight parameters (i.e., the *z* matrices in the equations) were tuned in a simulated "babbling phase". In this phase, the cells specifying the motor cortex movement command (*M*) were randomly activated in a time-varying manner, leading to time-varying articulator movements (*Artic*) and an accompanying acoustic signal (*Acoust*). The motor commands *M* were used in combination with the resulting auditory (*A*) and somatosensory (*S*) feedback to tune the synaptic weight matrices z_{AuM} and z_{SM} (see Appendix B for details regarding the algorithms used to tune the model's synaptic weights).

After the babbling phase, the model was trained to produce a small corpus of speech sounds (consisting of individual phonemes, syllables, and short words) via a process meant to approximate an infant learning a new sound by hearing it from an adult and then trying to produce it a few times. For each sound, the model was first presented with an acoustic example of the sound while simultaneously activating a speech sound map cell (*P*) that was chosen to represent the new sound. The resulting spatiotemporal auditory pattern (*A)* was used to tune the synaptic weights representing the auditory target for the sound (z_{PAu}). Then a short "practice phase", involving approximately 5–10 attempts to produce the sound by the model, was used to tune the synaptic weights making up the feedforward commands for the sound (z_{PM}). Finally, after the feedforward weights were tuned, additional repetitions were used to tune the somatosensory target for the sound (z_{PS}).

Simulation 1: "good doggie"

For this simulation, an utterance of the phrase "good doggie" was recorded at a sampling rate of 10 kHz. Formants were extracted from the signal and were modified slightly to form an auditory target that better matched the vocal tract characteristics of the Maeda synthesizer. The auditory target was represented as a convex region for each time point (see Guenther, 1998 for a discussion of convex region targets). Figure 3 shows the results of the simulations through the spectrograms of model utterances. The top plot shows the original spectrogram. The remaining plots show the 1st, 3rd, 5th, 7th, and 9th model attempts to produce the sound. With each trial, the feedforward system subsumes the corrective commands generated by the feedback system to compensate for the sensory error signals that arise during that trial. As can be seen from the figure, the spectrograms approach the original as learning progresses.

Simulation 2: Abbs and Gracco (1984) lip perturbation

In this simulation of the lip perturbation study, the model's lower lip was perturbed downward using a steady force during the movement toward closure of the lips when producing the utterance /aba/. Figure 4 shows a comparison of the model's productions to those measured in the original experiment for normal (no perturbation) and perturbed trials. The experiment results demonstrated that the speech motor system compensates for the perturbation by lowering the upper lip further than normal, resulting in successful closure of the lips despite the downward perturbation to the lower lip. The corresponding model simulations are shown in the right panel of Figure 4. The model was first trained to produce the utterance /aba/. After the sound was learned, the lower lip parameter of the model was perturbed with a constant downward force. The onset of perturbation was determined by tracking the velocity of the jaw parameter. The vertical black line marks the onset of perturbation. The position of the lips during the control condition is shown with the dashed lines while the position during the perturbed condition is shown with the solid lines. When the lips are perturbed, the tactile and proprioceptive feedback no longer matches the somatosensory target, giving rise to a somatosensory error signal and corrective motor command through the model's feedback subsystem. The commandis generated approximately 60 ms (the sum of τ_{ArS}, τ_{SM}, and τ_{MAr}) after the onset of perturbation. This is within the range of values (22–75 ms) measured during the experiment.

Simulation 3: Kelso, Tuller, Vatikiotis-Bateson, and Fowler (1984) jaw perturbation

In the experiment, the jaw was perturbed downward during the upward movement of the closing gesture in each of the two words: /baeb/ and /baez/. Their results demonstrate that the upper lip compensated for the perturbation during the production of /baeb/ but not during the production of /baez/ (top panel of Figure 5). These results indicate that compensation to perturbation does not affect the whole vocal tract but primarily affects articulators involved in the production of the particular phonetic unit that was being perturbed. Since the upper lip is not involved in the production of /z/, it is not influenced by the jaw perturbation in /baez/.

In the model simulations (bottom panel of Figure 5), we used the words /baeb/ and /baed/ to demonstrate the effects of jaw perturbation[10]. A steady perturbation corresponding to the increased load in the experiments was applied during the upward movement of the jaw. The perturbation was simulated by adding a constant value to the jaw height articulator of the vocal tract model. The perturbation remained in effect through the end of the utterance, as in the experiment. The onset of the perturbation is indicated by the vertical line in the simulation diagrams of Figure 5 and was determined by the velocity and position of the jaw displacement.

[10]The model is currently not capable of producing fricatives such as /z/, so instead the phoneme /d/, which like /z/ involves an alveolar constriction of the tongue rather than a lip constriction, was used.

The dotted lines indicate the positions of the articulators in the normal (unperturbed) condition. The solid lines indicate the positions in the perturbed condition. As in the experiment, the upper lip compensates by moving further downward when the bilabial stop /baeb/ is perturbed, but not when the alveolar stop /baed/ is perturbed.

COMPARING THE MODEL'S CELL ACTIVITIES TO THE RESULTS OF FMRI STUDIES

As stated in the Introduction, a major goal of the current modeling work is to provide a framework for interpreting the results of neuroimaging studies of speech production, and for generating predictions to help guide future neuroimaging studies. To this end, we have identified likely neuroanatomical locations of the model's components based on the results of previous neurophysiological studies as well as the results of functional magnetic resonance imaging experiments conducted by our laboratory. These locations allow us to run "simulated fMRI experiments" in which the model produces speech sounds in different speaking conditions, and the model cell activities are then used to generate a simulated hemodynamic response pattern based on these cell activations. These simulated hemodynamic response patterns can then be compared to the results of fMRI and/or positron emission tomography (PET) experiments in which human subjects produce the same (or similar) speech sounds in the same speaking conditions. In this section we describe this simulation process and the resulting hemodynamic response patterns, including a comparison of these patterns to the results of an fMRI experiment of simple syllable production performed in our laboratory. The results in this section are meant to illustrate the degree to which the model can currently account for the brain activities seen in human speech production experiments, and to serve as a baseline for future simulations involving additional speaking conditions that will test specific hypotheses generated from the model.

In Appendix A we detail the hypothesized anatomical locations of the model's components, with particular reference to the brain of the canonical single subject provided with the SPM2 software package (Friston, Ashburner, Holmes, & Poline, 2002). These locations are given in Montreal Neurological Institute (MNI) normalized spatial coordinates in addition to anatomical descriptions with reference to specific sulci and gyri. Figure 6 illustrates the locations of the model's components projected onto the lateral surface of the standard SPM brain, with the corresponding MNI coordinates provided in Table 1 of Appendix A.

FMRI and PET studies of speech production typically involve one or more "speaking conditions", in which the subject produces speech, and a "baseline condition", in which the subject rests quietly. The brain regions that become "active" during speech (i.e., those that have a larger hemodynamic response in the speech condition compared to the baseline condition) are typically interpreted as being involved in speech production.

In the model simulations, the "speaking condition" consisted of the model producing simple syllables. That is, speech sound map cells corresponding to the syllables were activated (Equation 1), and Equations 2–10 were used to calculate the activities of the model's cells (with the same model parameters used in the jaw and lip perturbation simulations described above). In the "baseline condition" all model cell activities were set to zero, corresponding to a resting state in which no speech is being produced. To produce the simulated hemodynamic response for each condition, model cell activities were first normalized by the maximum possible activity of the cell; this was done to correct for differences in the dynamic ranges of the different cell types in the model. The resultant activity was then convolved with an idealized hemodynamic response function, generated using default settings of the function 'spm_hrf' from the SPM toolbox. This function was designed by the creators of SPM to approximate the transformation from cell activity to hemodynamic response in the brain. For brain locations that include more

than one cell at the same location (i.e., those with the same MNI coordinates in Table 1 of Appendix A) the overall hemodynamic response was simply the sum of the responses of the individual cells. A brain volume was then constructed with the appropriate hemodynamic response values at each position. Responses were smoothed with a Gaussian kernel (FWHM=12mm) to approximate the smoothing carried out during standard SPM analysis of human subject data[11]. The resultant volume was then rendered using routines from the SPM toolbox.

In order to qualitatively compare the model's simulated activations with those of actual speakers, we conducted an fMRI experiment in which ten subjects produced simple consonant-vowel (CV) syllables that were read from a display screen in the scanner. Blood oxygenation level dependent (BOLD) responses were collected in 10 neurologically normal, right-handed speakers of American English (3 female, 7 male) during spoken production of vowel-vowel (VV), consonant-vowel (CV), and CVCV syllables which were presented visually (spelled out, e.g. "pah"). An event-triggered paradigm with a 15–18 second interstimulus interval was used wherein two whole head functional scans (3 seconds each in duration) were collected shortly after each syllable production, timed to occur near the peak of the speech-related hemodynamic response (approximately 4–6 seconds after the syllable is spoken). Since no scanning was done while the subject was pronouncing a syllable, this paradigm avoids confounds due to scanner noise during speech as well as image artifacts due to articulator motion. One to three runs of approximately 20 minutes each were completed for each subject. Data were obtained using a whole head coil in Siemens Allegra (6 Subjects) and Trio (4 subjects) scanners. Thirty axial slices (5 mm thick, 0 mm skip) parallel to the anterior and posterior commissure covering the whole brain were imaged with a temporal resolution of 3 sec using a T2*-weighted pulse sequence (TR=3s, TE=30ms, flip angle=90°, FOV=200mm and interleaved scanning). Images were reconstructed as a 64 × 64 × 30 matrix with a spatial resolution of 3.1×3.1×5 mm. To aid in the localization of functional data and for generating regions of interest (ROIs), high-resolution T1-weighted 3D MRI data were collected with the following parameters: TR=6.6ms, TE=2.9ms, flip angle=8°, 128 slices in sagittal place, FOV=256mm. Images were reconstructed as a 256 × 256 × 128 matrix with a 1 × 1 × 1.33 mm spatial resolution. The data from each subject were corrected for head movement, coregistered with the high-resolution structural image and normalized to MNI space. Random effects analysis was performed on the data using the SPM toolbox. The results were thresholded using a false discovery rate of $p<0.05$ (corrected).

Brain activations during syllable production (as compared to a baseline task involving passive viewing of visually presented X's on the display) are shown in the left half of Figure 7. The right half of Figure 7 shows brain activations derived from the DIVA model while producing the same syllables, with the model's components localized on the cortical surface and cerebellum as described in Appendix A. For the most part, the model's activations are qualitatively similar to those of the fMRI subjects. The biggest difference in activation concerns the supplementary motor area in the medial frontal lobe. This area, which is active in the experimental subjects but is not included in the mode at this time, is believed to be involved in the initiation and/or sequencing of speech sounds (see Concluding Remarks for details). Another difference concerns the respiratory portion of the motor cortex, high up on the motor strip, which is more active in the model than in the experimental subjects. This may be due to the fact that the model has no activity in this area during the baseline condition (quiet resting), whereas experimental subjects continue breathing during the baseline condition, perhaps controlled in part by motor cortex. The reduced baseline respiratory motor cortex activity in

[11]Each of the model's cells is treated as occupying a single point in MNI space. However we believe that each model cell corresponds to a small population of neurons, rather than a single neuron, distributed across a small portion of cortex. A second purpose of the Gaussian smoothing is to approximate this population distribution.

the model would result in greater activity for the model than for subjects in the speech – baseline comparison.

Although it is informative to see how much of the fMRI activity in human subjects producing simple syllables can be accounted for by the model, it is perhaps more informative to generate novel predictions from the model and test them in future neuroimaging studies. We are currently performing two such fMRI studies, one involving somatosensory perturbation during speech (using a pneumatic bite block) and one involving real-time auditory perturbation of the subject's acoustic feedback of their own speech. According to the model, somatosensory perturbation should lead to activity of somatosensory error cells in the anterior supramarginal gyrus (ΔS in Figure 6) due to a mismatch between the somatosensory target region and the incoming somatosensory feedback. Such activity would not be expected during unperturbed speech since the feedforward command in adults is well-tuned and thus few if any somatosensory errors should arise without perturbation. Similarly, auditory perturbation during speech should lead to more activation of auditory error cells in the superior temporal gyrus and planum temporale (ΔA in Figure 6) than unperturbed speech. The results of these fMRI studies should help us further refine our account of the neural bases of speech production. We also plan to investigate quantitative techniques for comparing model and human activations. One possible measure is *mutual information* (e.g., Maes et al., 1997), which describes the degree of agreement between two datasets in a way that is more robust than other comparable measures such as correlation.

CONCLUDING REMARKS

In this article we have described a neural model that provides a unified account for a wide range of speech acoustic, kinematic, and neuroimaging data. New computer simulations of the model were presented to illustrate the model's ability to provide a detailed account for experiments involving compensations to perturbations of the lip and jaw. With the goal of providing a computational framework for interpreting functional neuroimaging data, we have explicitly identified expected anatomical locations of the model's components, and we have compared the model's activities to activity measured using fMRI during simple syllable production and with and without a jaw perturbation.

Although the model described herein accounts for most of the activity seen in fMRI studies of speech production, it does not provide a complete account of the cortical and cerebellar mechanisms involved. In particular, as currently defined, the DIVA model is given a phoneme string by the modeler, and the model produces this phoneme string in the specified order. Brain structures involved in the selection, initiation, and sequencing of speech movements are not treated in the preceding discussion; these include the anterior cingulate area, the supplementary motor area (SMA), the basal ganglia, and (possibly) the anterior insula. The anterior cingulate gyrus lies adjacent to the SMA on the medial surface of the cortex in the interhemispheric fissure. This area is known to be involved in initiation of self-motivated behavior. Bilateral damage to the anterior cingulate area can result in akinetic mutism, characterized by a profound inability to initiate movements (DeLong, 1999). The anterior cingulate has also been implicated in execution of appropriate verbal responses and suppression of inappropriate responses (Paus, Petrides, Evans, & Meyer, 1993; Buckner et al., 1996; Nathaniel-James, Fletcher, & Frith, 1997). Several researchers have posited that the supplementary motor area is particularly involved for self-initiated responses, i.e., responses made in the absence of external sensory cues, whereas lateral premotor cortex is more involved when responding to external cues (e.g., Goldberg, 1985; Passingham, 1993). As the model is currently defined, it is not possible to differentiate between internally generated and externally cued speech. Diseases of the basal ganglia, such as Parkinson's disease and Huntington's disease, are known to impair movement sequencing (Stern, Mayeux, Rosen, & Ilson, 1983; Georgiou et al., 1994; Phillips, Chiu,

Bradshaw, & Iansek, 1995; Rogers, Phillips, Bradshaw, Iansek, & Jones, 1998), and single-cell recordings indicate that cells in the basal ganglia in monkeys and rats code aspects of movement sequences (Kermadi & Joseph, 1995; Aldridge & Berridge, 1998). The basal ganglia are strongly interconnected to the frontal cortex through a set of segregated basal ganglia-thalamo-cortical loops, including a loop focused on the SMA (DeLong & Wichman, 1993; Redgrave, Prescott, & Gurney, 1999). Like the SMA, the basal ganglia appear to be especially important when movements must be selected and initiated in the absence of external cues (Georgiou et al., 1994; Rogers et al., 1998). Also, stimulation at the thalamic stage of the basal ganglia-thalamo-cortical loops has been shown to affect the rate of speech production (Mateer, 1978). The lesion study of Dronkers (1996) indicated that the anterior insular cortex, or insula, buried in the sylvian fissure near the base of premotor cortex, plays an important role in speech production since damage to the insula is the likely source of pure apraxia of speech, a disorder involving an inability to select the appropriate motor programs for speech. Others have identified insula activation in certain speech tasks (e.g., Wise et al., 1999; Nota & Honda, 2003). The fMRI study of Nota and Honda (2003) suggests that the insula becomes involved when different syllables have to be sequenced in a particular order, as opposed to repetitive production of the same syllable. Based on these studies, we hypothesize that the insula plays a role in selecting the proper speech sound map cells in the ventral lateral premotor cortex.

Some additional factors limit the biological plausibility of the model in its current form. First, as described herein, all model cells of a particular type (e.g., the motor position cells) typically become active simultaneously. However, studies of primate cortex typically identify "recruitment curves" that show a more gradual onset of cells in a particular brain region (e.g., Kalaska & Crammond, 1992). Second, we make a sharp distinction between premotor cortex and primary motor cortex, with premotor cortex involving higher-level representations (the speech sound map) and motor cortex involving low-level motor representations (the articulator velocity and position cells). Neurophysiological results indicate that, instead, there appears to be a continuum of cells from motor to premotor cortex, with the complexity of the motor representation increasing as one moves anteriorly along the precentral gyrus into the premotor cortex (e.g., Kalaska et al., 1992). Future work will involve modifications that make the model more compatible with these findings.

Finally, it is interesting to note that the current model provides a more detailed account of the "mental syllabary" concept described by Levelt and colleagues (e.g., Levelt et al., 1994). In our account, the speech sound map cells can be thought of as the primary component of the syllabary, but additional components include the feedforward command pathways to motor cortex (the "gestural score"), and the auditory and somatosensory target projections to the higher-order auditory and somatosensory cortices. Thus in our view the syllabary is best thought of as a network of regions that together constitute the sensorimotor representation of frequently produced syllables.

Acknowledgements

Supported by the National Institute on Deafness and other Communication Disorders (R01 DC02852, F. Guenther PI; S. Ghosh also supported by R01 DC01925, J. Perkell PI).

References

Abbs JH, Gracco VL. Control of complex motor gestures: orofacial muscle responses to load perturbations of lip during speech. Journal of Neurophysiology 1984;51:705–723. [PubMed: 6716120]

Ackermann H, Vogel M, Petersen D, Poremba M. Speech deficits in ischaemic cerebellar lesions. Journal of Neurology 1992;239:223–227. [PubMed: 1597689]

Aldridge JW, Berridge KC. Coding of serial order by neostriatal neurons: a "natural action" approach to movement sequence. Journal of Neuroscience 1998;18:2777–2787. [PubMed: 9502834]

Andersen RA. Multimodal integration for the representation of space in the posterior parietal cortex. *Philosophical Transactions of the Royal Society of London.* Series B: Biological Sciences 1997;352:1421–1428.

Arbib, M.A. (in press). From monkey-like action recognition to human language: An evolutionary framework for linguistics. *Behavioral and Brain Sciences*

Binkofski F, Buccino G. Motor functions of the Broca's region. Brain and Language 2004;89:362–369. [PubMed: 15068919]

Boling W, Reutens DC, Olivier A. Functional topography of the low postcentral area. Journal of Neurosurgery 2002;97:388–395. [PubMed: 12186467]

Browman CP, Goldstein L. Articulatory gestures as phonological units. Phonology 1989;6:201–251.

Buchsbaum BR, Hickok G, Humphries C. Role of left posterior superior temporal gyrus in phonological processing for speech perception and production. Cognitive Science 2001;25:663–678.

Buckner RL, Raichle ME, Miezin FM, Petersen SE. Functional anatomic studies of memory retrieval for auditory words and visual pictures. Journal of Neuroscience 1996;16:6219–6235. [PubMed: 8815903]

Callan DE, Kent RD, Guenther FH, Vorperian HK. 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 2000;43:721–736.

Caplan D, Gow D, Makris N. Analysis of Lesions by Mri in Stroke Patients with Acoustic-Phonetic Processing Deficits. Neurology 1995;45:293–298. [PubMed: 7854528]

Celsis P, Boulanouar K, Doyon B, Ranjeva JP, Berry I, Nespoulous JL, et al. Differential fMRI responses in the left posterior superior temporal gyrus and left supramarginal gyrus to habituation and change detection in syllables and tones. Neuroimage 1999;9:135–144. [PubMed: 9918735]

Corfield DR, Murphy K, Josephs O, Fink GR, Frackowiak RS, Guz A, et al. Cortical and subcortical control of tongue movement in humans: a functional neuroimaging study using fMRI. Journal of Applied Physiology 1999;86:1468–1477. [PubMed: 10233106]

Creutzfeldt O, Ojemann G, Lettich E. Neuronal-Activity in the Human Lateral Temporal-Lobe.1. Responses to Speech. Experimental Brain Research 1989a;77:451–475.

Creutzfeldt O, Ojemann G, Lettich E. Neuronal-Activity in the Human Lateral Temporal-Lobe.2. Responses to the Subjects Own Voice. Experimental Brain Research 1989b;77:476–489.

Damasio H, Damasio AR. The anatomical basis of conduction aphasia. Brain 1980;103:337–350. [PubMed: 7397481]

DeLong, M. R. (1999). The basal ganglia. In E.R. Kandel, J. H. Schwartz, & T. M. Jessell (Eds.), *Principles of Neural Science* (4th ed., pp. 853–867). New York: McGraw Hill.

DeLong MR, Wichman T. Basal ganglia-thalamocortical circuits in Parkinsonian signs. Clinical Neuroscience 1993;1:18–26.

di Pellegrino G, Fadiga L, Fogassi L, Gallese V, Rizzolatti G. Understanding motor events: a neurophysiological study. Experimental Brain Research 1992;91:176–180.

Dronkers NF. A new brain region for coordinating speech articulation. Nature 1996;384:159–161. [PubMed: 8906789]

Eliades SJ, Wang X. Sensory-motor interaction in the primate auditory cortex during self-initiated vocalizations. J Neurophysiol 2003;89:2194–2207. [PubMed: 12612021]

Evans KC, Shea SA, Saykin AJ. Functional MRI localisation of central nervous system regions associated with volitional inspiration in humans. Journal of Physiology 1999;520(Pt 2):383–392. [PubMed: 10523407]

Ferrari PF, Gallese V, Rizzolatti G, Fogassi L. Mirror neurons responding to the observation of ingestive and communicative mouth actions in the monkey ventral premotor cortex. Eur J Neurosci 2003;17:1703–1714. [PubMed: 12752388]

Fesl G, Moriggl B, Schmid UD, Naidich TP, Herholz K, Yousry TA. Inferior central sulcus: variations of anatomy and function on the example of the motor tongue area. Neuroimage 2003;20:601–610. [PubMed: 14527621]

Fink GR, Corfield DR, Murphy K, Kobayashi I, Dettmers C, Adams L, et al. Human cerebral activity with increasing inspiratory force: a study using positron emission tomography. Journal of Applied Physiology 1996;81:1295–1305. [PubMed: 8889766]

Foerster O. The motor cortex of man in the light of Hughlings Jackson's doctrines. Brain 1936;59:135–159.

Fox PT, Huang A, Parsons LM, Xiong JH, Zamarippa F, Rainey L, et al. Location-probability profiles for the mouth region of human primary motor-sensory cortex: model and validation. Neuroimage 2001;13:196–209. [PubMed: 11133322]

Friston, K. J., Ashburner, J., Holmes, A., & Poline, J. (2002). Statistical Parameteric Mapping [http://www.fil.ion.ucl.ac.uk/spm/] [Computer software].

Fujimura, O. & Lovins, J. (1978). Syllables as concatenative phonetic units. In A. Bell & J. B. Hooper (Eds.), *Syllables and Segments* (pp. 107–120). Amsterdam: North Holland.

Fulton, J. F. (1938). *Physiology of the nervous system* London: Oxford University Press.

Georgiou N, Bradshaw JL, Iansek R, Phillips JG, Mattingley JB, Bradshaw JA. Reduction in external cues and movement sequencing in Parkinson's disease. Journal of Neurology, Neurosurgery and Psychiatry 1994;57:368–370.

Geschwind N. Disconnexion syndromes in animals and man. II. Brain 1965;88:585–644. [PubMed: 5318824]

Goldberg G. Supplementary motor area structure and function: Review and hypotheses. Behavioral Brain Research 1985;8:567–588.

Graziano MS, Taylor CS, Moore T, Cooke DF. The cortical control of movement revisited. Neuron 2002;36:349–362. [PubMed: 12408840]

Guenther FH. A neural network model of speech acquisition and motor equivalent speech production. Biological Cybernetics 1994;72:43–53. [PubMed: 7880914]

Guenther FH. Speech sound acquisition, coarticulation, and rate effects in a neural network model of speech production. Psychological Review 1995;102:594–621. [PubMed: 7624456]

Guenther, F. H. (1998). A theoretical framework for speech acquisition and production. In *Proceedings of the Second International Conference on Cognitive and Neural Systems* (pp. 57). Boston: Boston University Center for Adaptive Systems.

Guenther, F. H. (2001). Neural modeling of speech production. In *Proceedings of the 4th International Nijmegen Speech Motor Conference*

Guenther FH, Espy-Wilson CY, Boyce SE, Matthies ML, Zandipour M, Perkell JS. Articulatory tradeoffs reduce acoustic variability during American English /r/ production. Journal of the Acoustical Society of America 1999;105:2854–2865. [PubMed: 10335635]

Guenther, F. H. & Ghosh, S. S. (2003). A model of cortical and cerebellar function in speech. In *Proceedings of the XVth International Congress of Phonetic Sciences* (pp. 169–173).

Guenther FH, Hampson M, Johnson D. A theoretical investigation of reference frames for the planning of speech movements. Psychological Review 1998;105:611–633. [PubMed: 9830375]

Guenther FH, Nieto-Castanon A, Ghosh SS, Tourville JA. Representation of sound categories in auditory cortical maps. Journal of Speech, Language, and Hearing Research 2004;47:46–57.

Hashimoto Y, Sakai KL. Brain activations during conscious self-monitoring of speech production with delayed auditory feedback: an fMRI study. Human Brain Mapping 2003;20:22–28. [PubMed: 12953303]

Hickok G, Poeppel D. Dorsal and ventral streams: a framework for understanding aspects of the functional anatomy of language. Cognition 2004;92:67–99. [PubMed: 15037127]

Houde JF, Jordan MI. Sensorimotor adaptation in speech production. Science 1998;279:1213–1216. [PubMed: 9469813]

Houde JF, Nagarajan SS, Sekihara K, Merzenich MM. Modulation of the auditory cortex during speech: an MEG study. Journal of Cognitive Neuroscience 2002;14:1125–1138. [PubMed: 12495520]

Iacoboni M, Woods RP, Brass M, Bekkering H, Mazziotta JC, Rizzolatti G. Cortical mechanisms of human imitation. Science 1999;286:2526–2528. [PubMed: 10617472]

Indefrey P, Levelt WJM. The spatial and temporal signatures of word production components. Cognition 2004;92:101–144. [PubMed: 15037128]

Kalaska JF, Cohen DA, Hyde ML, Prud'homme M. A comparison of movement direction-related versus load direction-related activity in primate motor cortex, using a two-dimensional reaching task. Journal of Neuroscience 1989;9:2080–2102. [PubMed: 2723767]

Kalaska JF, Crammond DJ. Cerebral cortical mechanisms of reaching movements. Science 1992;255:1517–1523. [PubMed: 1549781]

Kawato M, Gomi H. A computational model of four regions of the cerebellum based on feedback-error learning. Biological Cybernetics 1992;68:95–103. [PubMed: 1486143]

Kelso JA, Tuller B, Vatikiotis-Bateson E, Fowler CA. Functionally specific articulatory cooperation following jaw perturbations during speech: evidence for coordinative structures. Journal of Experimental Psychology: Human Perception and Performance 1984;10:812–832. [PubMed: 6239907]

Kermadi I, Joseph JP. Activity in the caudate nucleus of monkey during spatial sequencing. Journal of Neurophysiology 1995;74:911–933. [PubMed: 7500161]

Kohler E, Keysers C, Umilta MA, Fogassi L, Gallese V, Rizzolatti G. Hearing sounds, understanding actions: action representation in mirror neurons. Science 2002;297:846–848. [PubMed: 12161656]

Krakauer, J. & Ghez, C. (1999). Voluntary movement. In E.R. Kandel, J. H. Schwartz, & T. M. Jessell (Eds.), *Principles of Neural Science* (4th ed., pp. 756–781). New York: McGraw Hill.

Levelt WJ, Roelofs A, Meyer AS. A theory of lexical access in speech production. Behavioral and Brain Sciences 1999;22:1–38. [PubMed: 11301520]

Levelt WJ, Wheeldon L. Do speakers have access to a mental syllabary? Cognition 1994;50:239–269. [PubMed: 8039363]

Lotze M, Erb M, Flor H, Huelsmann E, Godde B, Grodd W. fMRI evaluation of somatotopic representation in human primary motor cortex. Neuroimage 2000;11:473–481. [PubMed: 10806033]

Lotze M, Seggewies G, Erb M, Grodd W, Birbaumer N. The representation of articulation in the primary sensorimotor cortex. Neuroreport 2000;11:2985–2989. [PubMed: 11006980]

Luppino G, Murata A, Govoni P, Matelli M. Largely segregated parietofrontal connections linking rostral intraparietal cortex (areas AIP and VIP) and the ventral premotor cortex (areas F5 and F4). Experimental Brain Research 1999;128:181–187.

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

Maes F, Collignon A, Vandermeulen D, Marchal G, Seutens P. Multimodality image registration by maximisation of mutual information. IEEE Transactions on Medical Imaging 1997;16:187–197. [PubMed: 9101328]

Mateer C. Asymmetric effects of thalamic stimulation on rate of speech. Neuropsychologia 1978;16:497–499. [PubMed: 358009]

Matsumoto R, Nair DR, LaPresto E, Najm I, Bingaman W, Shibasaki H, et al. Functional connectivity in the human language system: a cortico-cortical evoked potential study. Brain 2004;127:2316–2330. [PubMed: 15269116]

Max L, Guenther FH, Gracco VL, Ghosh SS, Wallace ME. Unstable or insufficiently activated internal models and feedback-biased motor control as sources of dysfluency: A theoretical model of stuttering. Contemporary Issues in Communication Science and Disorders 2004;31:105–122.

McCarthy G, Allison T. Trigeminal evoked potentials in somatosensory cortex of the Macaca mulatta. Journal of Neurosurgery 1995;82:1015–1020. [PubMed: 7760175]

McCarthy G, Allison T, Spencer DD. Localization of the face area of human sensorimotor cortex by intracranial recording of somatosensory evoked potentials. Journal of Neurosurgery 1993;79:874–884. [PubMed: 8246056]

Meyer BU, Werhahn K, Rothwell JC, Roericht S, Fauth C. Functional organisation of corticonuclear pathways to motoneurones of lower facial muscles in man. Experimental Brain Research 1994;101:465–472.

Middleton FA, Strick PL. Cerebellar output channels. International Review of Neurobiology 1997;41:61–82. [PubMed: 9378611]

Miller JD. Auditory-perceptual interpretation of the vowel. Journal of the Acoustical Society of America 1989;85:2114–2134. [PubMed: 2659639]

Morasso, P., Sanguineti, V., & Frisone, F. (2001). Cortical maps as topology-representing neural networks applied to motor control: Articulatory speech synthesis. In K.A.H. Masterbroek & J. E. Vos (Eds.), *Plausible neural networks for biological modelling* (Dordrecht: Kluwer.

Morosan P, Rademacher J, Schleicher A, Amunts K, Schormann T, Zilles K. Human primary auditory cortex: cytoarchitectonic subdivisions and mapping into a spatial reference system. Neuroimage 2001;13:684–701. [PubMed: 11305897]

Nakamura A, Yamada T, Goto A, Kato T, Ito K, Abe Y, et al. Somatosensory homunculus as drawn by MEG. Neuroimage 1998;7:377–386. [PubMed: 9626677]

Nathaniel-James DA, Fletcher P, Frith CD. The functional anatomy of verbal initiation and suppression using the Hayling Test. Neuropsychologia 1997;35:559–566. [PubMed: 9106283]

Nieto-Castanon A, Ghosh SS, Tourville JA, Guenther FH. Region of interest based analysis of functional imaging data. Neuroimage 2003;19:1303–1316. [PubMed: 12948689]

Nieto-Castanon, A., Guenther, F.H., Perkell, J.S., and Curtin, H. (in press). A modeling investigation of articulatory variability and acoustic stability during American English /r/ production. *Journal of the Acoustical Society of America*

Nota, Y. & Honda, K. (2003). Possible role of the anterior insula in speech production. In *Proceedings of the Sixth International Seminar on Speech Production* (pp. 191–194). Macquarie Centre for Cognitive Science.

Numminen J, Curio G. Differential effects of overt, covert and replayed speech on vowel-evoked responses of the human auditory cortex. Neuroscience Letters 1999;272:29–32. [PubMed: 10507535]

Numminen J, Salmelin R, Hari R. Subject's own speech reduces reactivity of the human auditory cortex. Neuroscience Letters 1999;265:119–122. [PubMed: 10327183]

O'Brien JH, Pimpaneau A, Albe-Fessard D. Evoked cortical responses to vagal, laryngeal and facial afferents in monkeys under chloralose anaesthesia. Electroencephalography and Clinical Neurophysiology 1971;31:7–20. [PubMed: 4105847]

Olson JM, Marsh RL. Activation patterns and length changes in hindlimb muscles of the bullfrog Rana catesbeiana during jumping. Journal of Experimental Biology 1998;201(Pt 19):2763–2777. [PubMed: 9732331]

Passingham, R. E. (1993). *The frontal lobes and voluntary action* Oxford: Oxford University Press.

Paus T, Petrides M, Evans AC, Meyer E. Role of the human anterior cingulate cortex in the control of oculomotor, manual, and speech responses: a positron emission tomography study. Journal of Neurophysiology 1993;70:453–469. [PubMed: 8410148]

Penfield, W. & Rasmussen, T. (1950). *The cerebral cortex of mana clinical study of localization of function* New York: Macmillan.

Penfield, W. & Roberts, L. (1959). *Speech and brain-mechanisms* Princeton, N.J: Princeton University Press.

Perkell JS, Guenther FH, Lane H, Matthies ML, Perrier P, Vick J, et al. A theory of speech motor control and supporting data from speakers with normal hearing and profound hearing loss. Journal of Phonetics 2000;28:233–272.

Perkell, J. S., Matthies, M. L., Tiede, M., Lane, H., Zandipour, M., Marrone, N. et al. The distinctness of speakers' /s-sh/ contrast is related to their auditory discrimination and use of an articulatory saturation effect. *Journal of Speech, Language, and Hearing Research,* (in press).

Phillips JG, Chiu E, Bradshaw JL, Iansek R. Impaired movement sequencing in patients with Huntington's disease: a kinematic analysis. Neuropsychologia 1995;33:365–369. [PubMed: 7792003]

Picard C, Olivier A. Sensory cortical tongue representation in man. Journal of Neurosurgery 1983;59:781–789. [PubMed: 6604794]

Poeppel D. The analysis of speech in different temporal integration windows: Cerebral lateralization as 'asymmetric sampling in time'. Speech Communication 2003;41:245–255.

Redgrave P, Prescott TJ, Gurney K. The basal ganglia: a vertebrate solution to the selection problem? Neuroscience 1999;89:1009–1023. [PubMed: 10362291]

Riecker A, Ackermann H, Wildgruber D, Dogil G, Grodd W. Opposite hemispheric lateralization effects during speaking and singing at motor cortex, insula and cerebellum. Neuroreport 2000;11:1997–2000. [PubMed: 10884059]

Riecker A, Wildgruber D, Grodd W, Ackermann H. Reorganization of Speech Production at the Motor Cortex and Cerebellum following Capsular Infarction: a Follow-up Functional Magnetic Resonance Imaging Study. Neurocase 2002;8:417–423. [PubMed: 12529451]

Rivier F, Clarke S. Cytochrome oxidase, acetylcholinesterase, and NADPH-diaphorase staining in human supratemporal and insular cortex: evidence for multiple auditory areas. Neuroimage 1997;6:288–304. [PubMed: 9417972]

Rizzolatti G, Arbib MA. Language within our grasp . Trends in Neurosciences 1998;21:188–194. [PubMed: 9610880]

Rizzolatti G, Camarda R, Fogassi L, Gentilucci M, Luppino G, Matelli M. Functional organization of inferior area 6 in the macaque monkey. II. Area F5 and the control of distal movements. Experimental Brain Research 1988;71:491–507.

Rizzolatti G, Fadiga L, Gallese V, Fogassi L. Premotor cortex and the recognition of motor actions. *Brain Research.* Cognitive Brain Research 1996;3:131–141. [PubMed: 8713554]

Rizzolatti G, Fogassi L, Gallese V. Parietal cortex: from sight to action. Current Opinion in Neurobiology 1997;7:562–567. [PubMed: 9287198]

Rizzolatti G, Luppino G. The cortical motor system. Neuron 2001;31:889–901. [PubMed: 11580891]

Rogers MA, Phillips JG, Bradshaw JL, Iansek R, Jones D. Provision of external cues and movement sequencing in Parkinson's disease. Motor Control 1998;2:125–132. [PubMed: 9644283]

Saltzman EL, Munhall KG. A dynamical approach to gestural patterning in speech production. Ecological Psychology 1989;1:333–382.

Schmahmann JD, Pandya DN. The cerebrocerebellar system. International Review of Neurobiology 1997;41:31–60. [PubMed: 9378595]

Schroeder CE, Foxe JJ. The timing and laminar profile of converging inputs to multisensory areas of the macaque neocortex. *Brain Research.* Cognitive Brain Research 2002;14:187–198. [PubMed: 12063142]

Schwartz, J.-L. & Boë, L.-J. (2000). Predicting palatal contacts from jaw and tongue commands: a new sensory model and its potential use in speech control. In *Proceedings of the 5th Seminar On Speech Production: Models and Data* Institut fuer Phonetik und Sprachliche Kommunikation.

Simonyan K, Jurgens U. Cortico-cortical projections of the motorcortical larynx area in the rhesus monkey. Brain Research 2002;949:23–31. [PubMed: 12213296]

Stern Y, Mayeux R, Rosen J, Ilson J. Perceptual motor dysfunction in Parkinson's disease: a deficit in sequential and predictive voluntary movement. Journal of Neurology, Neurosurgery and Psychiatry 1983;46:145–151.

Tai YF, Scherfler C, Brooks DJ, Sawamoto N, Castiello U. The human premotor cortex is 'mirror' only for biological actions. Curr Biol 2004;14:117–120. [PubMed: 14738732]

Tallal, P., Miller, S., & Fitch, R.H. (1993). Neurobiological basis of speech: a case for the preeminence of temporal processing. In Tallal, P., Galaburda, A.M., Llinas, R., and von Euler, C. (Eds), *Temporal Information Processing in the Nervous System, with Special Reference to Dyslexia and Dysphasia* (pp. 27–47). New York: New York Academy of Sciences.

Tomaiuolo F, MacDonald JD, Caramanos Z, Posner G, Chiavaras M, Evans AC, et al. Morphology, morphometry and probability mapping of the pars opercularis of the inferior frontal gyrus: an in vivo MRI analysis. European Journal of Neuroscience 1999;11:3033–3046. [PubMed: 10510168]

Tremblay S, Shiller DM, Ostry DJ. Somatosensory basis of speech production. Nature 2003;423:866–869. [PubMed: 12815431]

Tzourio-Mazoyer N, Landeau B, Papathanassiou D, Crivello F, Etard O, Delcroix N, et al. Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. Neuroimage 2002;15:273–289. [PubMed: 11771995]

Urasaki E, Uematsu S, Gordon B, Lesser RP. Cortical tongue area studied by chronically implanted subdural electrodes--with special reference to parietal motor and frontal sensory responses. Brain 1994;117(Pt 1):117–132. [PubMed: 8149206]

Urban PP, Marx J, Hunsche S, Gawehn J, Vucurevic G, Wicht S, et al. Cerebellar speech representation: lesion topography in dysarthria as derived from cerebellar ischemia and functional magnetic resonance imaging. Archives of Neurology 2003;60:965–972. [PubMed: 12873853]

Westermann G, Reck ME. A new model of sensorimotor coupling in the development of speech. Brain and Language 2004;89:393–400. [PubMed: 15068923]

Wildgruber D, Ackermann H, Grodd W. Differential contributions of motor cortex, basal ganglia, and cerebellum to speech motor control: effects of syllable repetition rate evaluated by fMRI. Neuroimage 2001;13:101–109. [PubMed: 11133313]

Wise RJ, Greene J, Buchel C, Scott SK. Brain regions involved in articulation. Lancet 1999;353:1057–1061. [PubMed: 10199354]

Wise RJ, Scott SK, Blank SC, Mummery CJ, Murphy K, Warburton EA. Separate neural subsystems within 'Wernicke's area'. Brain 2001;124:83–95. [PubMed: 11133789]

Wise SP, Boussaoud D, Johnson PB, Caminiti R. Premotor and parietal cortex: corticocortical connectivity and combinatorial computations. Annual Review of Neuroscience 1997;20:25–42.

Zatorre RJ, Belin P, Penhune VB. Structure and function of auditory cortex: Music and speech. Trends in Cognitive Sciences 2002;6:37–46. [PubMed: 11849614]

Zatorre RJ, Evans AC, Meyer E, Gjedde A. Lateralization of phonetic and pitch discrimination in speech processing. Science 1992;256:846–849. [PubMed: 1589767]

APPENDIX A: ESTIMATED ANATOMICAL LOCATIONS OF THE MODEL'S COMPONENTS

In this appendix we describe hypothesized neuroanatomical locations of the model's components, including a treatment of the neurophysiological literature that was used to guide these location estimates. Table 1 summarizes the Montreal Neurological Institute (MNI) coordinates for each of the model's components; these coordinates were used to create the simulated fMRI activations shown in Figure 7. Unless otherwise noted, each cell type is represented symmetrically in both hemispheres. Currently there are no functional differences between the left and right hemisphere versions of a particular cell type in the model. However future versions of the model will incorporate hemispheric differences in cortical processing as indicated by experimental findings (e.g., Poeppel, 2003;Tallal et al., 1993;Zatorre et al., 1992,2002).

Motor Position and Velocity Maps

Cells coding for the position and velocity of the tongue parameters in the model are hypothesized to correspond with the Motor Tongue Area (MTA) as described by Fesl et al. (2003). The region lies along the posterior bank of the precentral gyrus roughly 2–3 cm above the Sylvian fissure. The spatial localization of this area is in agreement with imaging (Fesl et al., 2003; Corfield et al., 1999; Urasaki, Uematsu, Gordon, & Lesser, 1994; also see Fox et al., 2001) and physiological (Penfield & Rasmussen, 1950) studies of the primary motor region for tongue/mouth movements. We designated a motor (and somatosensory) tongue location for each degree of freedom in the model. This expanded representation is consistent with the large tongue sensorimotor representation

A region superior and medial to the tongue region along the posterior bank of the precentral gyrus has been shown to produce lip movements in humans when electrically stimulated (Penfield & Roberts, 1959). Comparing production of syllables involving tongue movements to those involving lip movements, Lotze et al. (2000b) found the lip area to be approximately 1–2 cm from the tongue area in the directions described by Penfield. In another mapping study of motor cortex using fMRI, Lotze et al. (2000a) showed the lip region inferolateral to the hand motor area, consistent with the Penfield electrical stimulation results. This area is hypothesized to code for the motor position and velocity of the model lip parameters. Upper and lower lip regions have been designated along the precentral gyrus superior and medial to the tongue representation. Data indicating the relative locations of upper and lower lip motor representations in humans is scarce. Currently, we have placed the upper lip motor

representation dorsomedial to the lower lip representation, mirroring the somatosensory organization (see Somatosensory State Map below).

Physiological recordings by Penfield and Roberts also indicate a primary motor region corresponding to jaw movements that lies between the lip and tongue representations along the posterior bank of the precentral sulcus, and a region corresponding to larynx motor control inferolateral to the tongue area (Penfield et al., 1959p. 200). Further evidence of the location of a motor larynx representation near the Sylvian fissure is provided by electrical stimulation in primates (e.g., Simonyan & Jurgens, 2002).

Fink et al. (1996) demonstrated dorsolateral precentral gyrus activation during voluntary breathing using PET. The bilateral region noted in that study lied along the superior portion of primary motor cortex, well above the ventral motor representations of the articulators. In an fMRI study, Evans, Shea, and Saykin (1999) found a similar activation association with volitional breathing along superior precentral gyrus medial to the Fink et al. findings and only in the left hemisphere. In the current study, we found activity in approximately the same regions as that described by Fink et al.: bilateral activation superior to and distinct from ventral motor activation (see left half of Figure 6). We hypothesize that this activity is associated with the control of breathing (e.g., maintenance of appropriate subglottal pressure) required for speech production and therefore place cells in this region that correspond to voicing control parameters in the model (specifically, parameter AGP of the Maeda articulatory synthesizer).

While the studies mentioned above indicate bilateral primary motor involvement during articulator movements, they do not explicitly show bilateral involvement of these areas during speech production (though Penfield & Roberts report a bilateral precentral gyrus region that causes “vocalization”). However, Indefrey and Levelt (2004), in their review of neuroimaging studies of speech note bilateral activation of ventral pre- and postcentral gyri during overt speech when compared to silence. In our fMRI results (left half of Figure 6) we found activation along both banks of the central sulcus in both hemispheres, but with stronger activation in the left hemisphere than the right. This finding is consistent with a report of bilateral primary motor activity during overt speech, but stronger activation in the left hemisphere (Riecker et al., 2000). In keeping with these findings, the model’s motor position and velocity cell populations are assumed to contain 20% more cells in the left hemisphere than the right hemisphere, resulting in the leftward bias of the model’s motor cortical activations in the right half of Figure 6.

We hypothesize that the model’s feedforward motor command (specifically, the product $P(t)$ $z_{PM}(t)$) involves a cerebellar contribution. Based on the lesion study by Ackermann, Vogel, Petersen, and Poremba (1992), the anterior paravermal region of the cerebellar cortex appears to play a role in the motor control of speech. A contribution to speech production by the medial anterior region of the cerebellum is also supported by a study of dysarthria lesions (Urban et al., 2003). Though not visible in Figure 6 because of the overlying cortex, our fMRI results also show superior medial cerebellum activation during CV production. Recent imaging studies (e.g., Riecker et al., 2000; Riecker, Wildgruber, Grodd, & Ackermann, 2002; Wildgruber et al., 2001) indicate bilateral cerebellum activation during speech production that lies posterior and lateral to the anterior paravermal activity. Our production results reveal distinct bilateral activations that lie behind the primary fissure and lateral to the cerebellum activity already mentioned, in roughly the same region described in these earlier studies. We have therefore placed model cells in two cerebellar cortical regions: anterior paravermal and superior lateral areas. Finally, we identify a region within the medial portion of the sub-cortical cerebellum where the deep cerebellar nuclei (the output cells of the cerebellum) are located.

Speech Sound Map

As described above, we believe the model's speech sound map consists of mirror neurons similar to those described by Rizzolatti and colleagues. Cells that behave in this fashion have been found in the left inferior premotor F5 region of the monkey (Rizzolatti et al., 1988; Rizzolatti et al., 1996a). Accordingly, we have designated a site in the left ventral premotor area, anterior to the precentral gyrus, as the speech sound map region. This is also consistent with our fMRI results (left half of Figure 6). The designated region, within ventral Brodmann's area 44 (the posterior portion of Broca's area), has been described as the human homologue of monkey area F5 (Rizzolatti & Arbib, 1998; Binkofski & Buccino, 2004)[12]. We expect that the speech sound map spreads into neighboring regions such as the precentral sulcus and anterior portion of the precentral gyrus.

Somatosensory State Map

Tactile and proprioceptive representations of the articulators are hypothesized to lie along the inferior postcentral gyrus, roughly adjacent to their motor counterparts across the central sulcus. Boling, Reutens, and Olivier (2002) demonstrated an anatomical marker for the tongue somatosensory region using PET imaging that built upon earlier work using electrical stimulation (Picard & Olivier, 1983). They describe the location of the tongue region below the anterior apex of the triangular region of the inferolateral postcentral gyrus approximately 2 cm above the Sylvian fissure. This region of the postcentral gyrus was found to represent the tongue in a somatosensory evoked potential study of humans (McCarthy, Allison, & Spencer, 1993), a finding further supported by a similar procedure in the macaque (McCarthy & Allison, 1995). By generating potentials on either side of the central sulcus, both studies by McCarthy and colleagues demonstrate adjacent motor-somatosensory organization of the tongue representation.

McCarthy et al. (1993) also mapped the primary sensory representations of the lip and palate. The lip representation was located superior and medial to the tongue representation along the anterior bank of the postcentral gyrus at the apex of the inferior postcentral triangle and below the hand representation. Nakamura et al. (1998) localized the lip and tongue sensory representations to nearly identical regions of the postcentral gyrus using MEG. The palatal representation was located inferolateral to the tongue region roughly 1 cm above the Sylvian fissure. The relative locations of the lip, tongue, and palate were confirmed in the macaque (McCarthy et al., 1995). Consistent with early electrophysiological work (Penfield et al., 1950) and a recent MEG study (Nakamura et al., 1998), we have placed the upper lip representation dorsomedial to the lower lip representation.

Graziano, Taylor, Moore, and Cooke (2002) report early electrical stimulation work (Fulton, 1938; Foerster, 1936) which depicts a sensory representation of the larynx at the inferior extent of the postcentral gyrus, near the Sylvian fissure. This location mirrors the motor larynx representation that lies on the inferior precentral gyrus.

Using the same reasoning as outlined above for the primary motor representation of articulators, we hypothesize bilateral somatosensory representations for each of the articulators, with a 20% leftward bias. As was the case for precentral activation, our fMRI results (Figure 6) show greater involvement of the left hemisphere postcentral gyrus.

[12]The rare bifurcation of the left ventral precentral sulcus (posterior segment intersects the central sulcus, anterior segment intersects the anterior ascending branch of the Sylvian fissure) on the SPM standard brain makes it difficult to localize ventral BA 44. No clear sulcal landmark distinguishes BA 44 from BA 6. We have placed the speech sound map region immediately behind the inferior end of the anterior ascending branch of the Sylvian fissure under the assumption that this area corresponds to ventral BA 44. The MNI coordinates chosen for the speech sound map are consistent with the inferior frontal gyrus pars opercularis region (Tzourio-Mazoyer et al., 2002).

Somatosensory Error Map

The DIVA model calls for the comparison of speech motor and somatosensory information for the purpose of somatosensory target learning and feedback-based control. We hypothesize that this component of the model, the somatosensory error map, lies within the inferior parietal cortex along the anterior supramarginal gyrus, posterior to the primary somatosensory representations of the speech articulators. Similarly, Hickok and colleagues (e.g., Hickok & Poeppel, 2004) have argued that speech motor commands and sensory feedback interface in the ventral parietal lobe, analogous to the visual-motor integration of the dorsal parietal lobe (Andersen, 1997; Rizzolatti, Fogassi, & Gallese, 1997). Reciprocal connections between area F5 and inferior parietal cortex has been demonstrated in the monkey (Luppino et al., 1999). These connections are believed to contribute to the sensorimotor transformations required to guide movements (see Rizzolatti & Lupppino, 2001) such as grasping. We hypothesize that similar connections are employed to monitor and guide speech articulator movements. Reciprocal connections between posterior inferior frontal gyrus and both the supramarginal gyrus and posterior superior temporal gyrus in the human have been demonstrated by Matsumoto et al. (2004) using a cortico-cortical evoked potential technique involving direct cortical stimulation in epilepsy patients.

Auditory State Map

The auditory state cells are hypothesized to lie within primary auditory cortex and the surrounding auditory association cortex. Therefore we have localized auditory state regions along the medial portion of Heschl's gyrus and the anterior planum temporale (Rivier & Clarke, 1997; Morosan et al., 2001). These locations are consistent with fMRI studies of speech perceptual processing performed by our group (Guenther, Nieto-Castanon, Ghosh, & Tourville, 2004).

Auditory Error Map

Hickok and colleagues have demonstrated an area within the left posterior Sylvian fissure at the junction of the temporal and parietal lobes (area SPT) and another in the lateral posterior superior temporal gyrus/sulcus that respond during speech perception and production (Buchsbaum et al., 2001; Hickok et al., 2004). The former area was also noted by Wise et al. (2001) in a review of several imaging studies of speech processing as being "engaged in the motor act of speech." Thus these areas could compare efferent motor commands with auditory input as in the model's auditory error map. Presently, insufficient information is available to differentiate between the two sites. Therefore we have placed auditory error cells at both locations.

The Buchsbaum and Hickok studies indicated that these regions might be lateralized to the left hemisphere. However, using delayed auditory feedback, Hashimoto and Sakai (2003) showed bilateral activation of the posterior superior temporal gyrus and the inferior supramarginal gyrus. Moreover, activity within the posterior superior temporal gyrus and superior temporal sulcus was correlated with size of the disfluency effect caused by the delayed auditory feedback. Based on this result, we have placed the auditory error cells bilaterally; however we consider it possible that these cells are left-lateralized, and further investigation of this issue is being carried out in ongoing studies of auditory and articulatory perturbation in our laboratory.

As mentioned above, Matsumoto et al., (2004) demonstrated bi-directional connections in humans between posterior inferior frontal gyrus and the two regions proposed to contain the speech error map. Evidence of modulation of the posterior superior temporal gyrus by speech

production areas in the human is also provided by the Wise et al. (1999) positron emission tomography study which demonstrated reduced superior temporal gyrus activation during a speech production task compared to a listening task. Single unit recordings from primate auditory cortex provide further support. Eliades & Wang (2002) noted suppression of marmoset auditory cortex immediately prior to self-initiated vocalizations. Based on these results we propose that projections from premotor to higher-order auditory cortical areas exist either directly or via an intermediate area (e.g., anterior supramarginal gyrus).

Although we have treated the auditory and somatosensory error maps as distinct entities in this discussion, we believe there probably exist combined somato-auditory cells, and somato-auditory error maps, that involve relatively highly processed combinations of speech-related somatosensory and auditory information. Thus we expect a continuum of sensory error map representations in and between the superior temporal gyrus, sylvian fissure, and supramarginal gyrus, rather than entirely distinct auditory and somatosensory error maps as described thus far.

APPENDIX B: TUNING THE SYNAPTIC WEIGHTS IN THE MODEL

For the simulations reported above, the model's synaptic weight parameters (i.e., the *z* matrices in the equations) were tuned as follows.

The synaptic weights z_{AuM} and z_{SM}, which encode the transformation from auditory (z_{AuM}) and somatosensory (z_{SM}) errors into corrective motor commands, were calculated by an explicit algorithm that determines the local pseudoinverse for any configuration of the vocal tract. A more biologically plausible method for tuning these weights was described in Guenther et al. (1998). The pseudoinverse ($\mathbf{J}(M)^{-1}$) is determined by applying a perturbation (δM) of the motor state (M), measuring the resulting change in sensory space ($\delta Sensory = [\delta S, \delta Au]$), and calculating the Jacobian and its inverse as follows:

$$\mathcal{J}(M) = \delta Sensory / \delta M$$
$$z_{AuM}(M) = \mathcal{J}(M)^{-1}.$$

The matrix z_{PAu} corresponds to the auditory expectation for a given speech sound target. This matrix was set to encode upper and lower bounds, for each 1 ms time slice, of the first three formants that were extracted from the acoustic sample.

The matrices z_{PM} and z_{PS}, which encode the feedforward command and somatosensory target for a sound, respectively, were updated during the practice phase. The matrix z_{PM} was updated using the feedback commands ($M_{Feedback}$) generated by the auditory portion of the feedback control subsystem, while the matrix z_{PS} was tuned based on the somatosensory error (ΔS). To account for temporal delays, these tuning processes align the auditory error or somatosensory error data slice with the appropriate time slices of the weight matrices. The weight update rules include this temporal alignment.

$$z_{PM}[t - \tau_{PAu}] = M_{Feedback}(t)$$
$$z_{PS}[t - \tau_{PS}] = \Delta S(t).$$

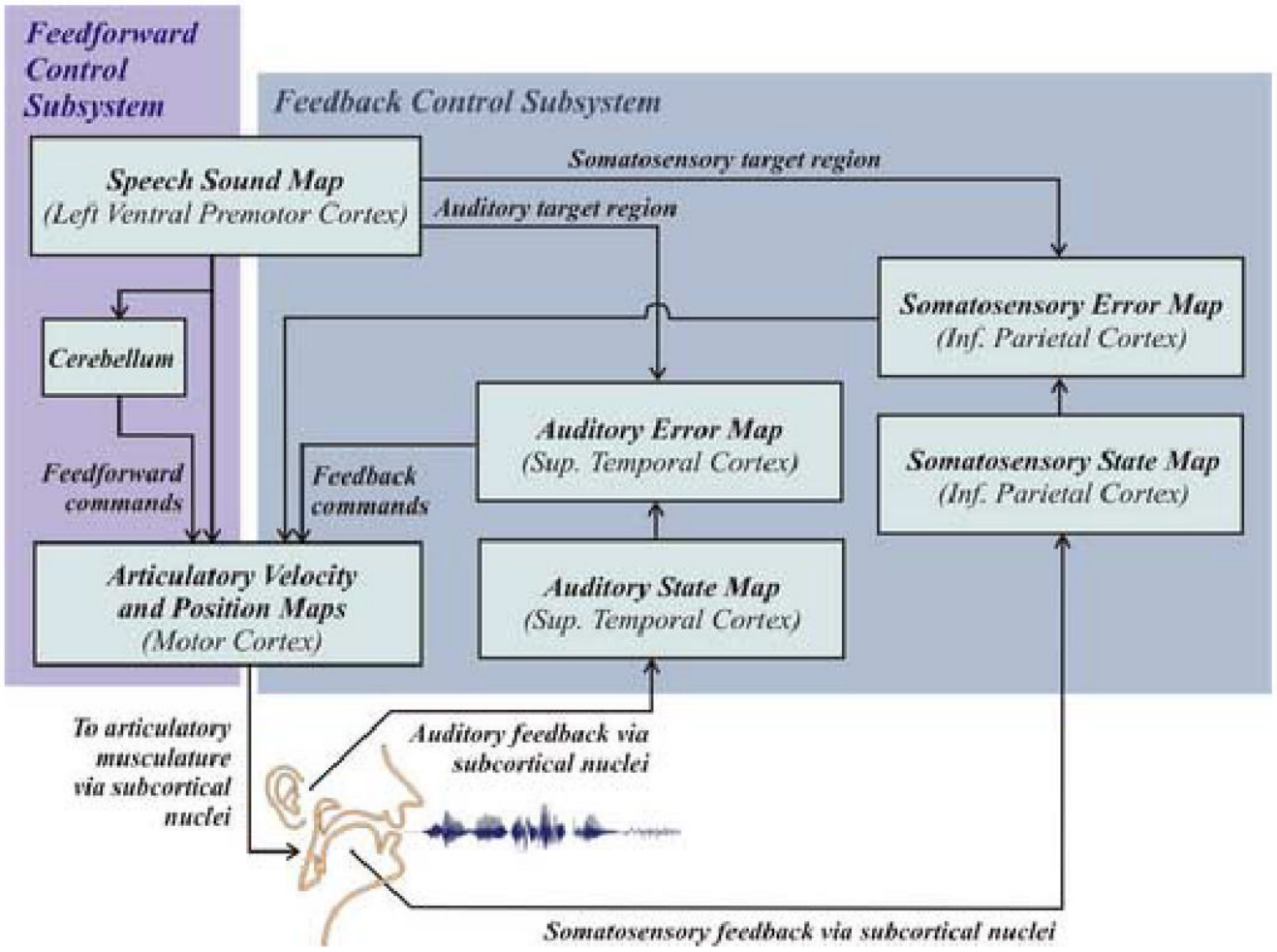

Figure 1.
Hypothesized neural processing stages involved in speech acquisition and production according to the DIVA model. Projections to and from the cerebellum are simplified for clarity.

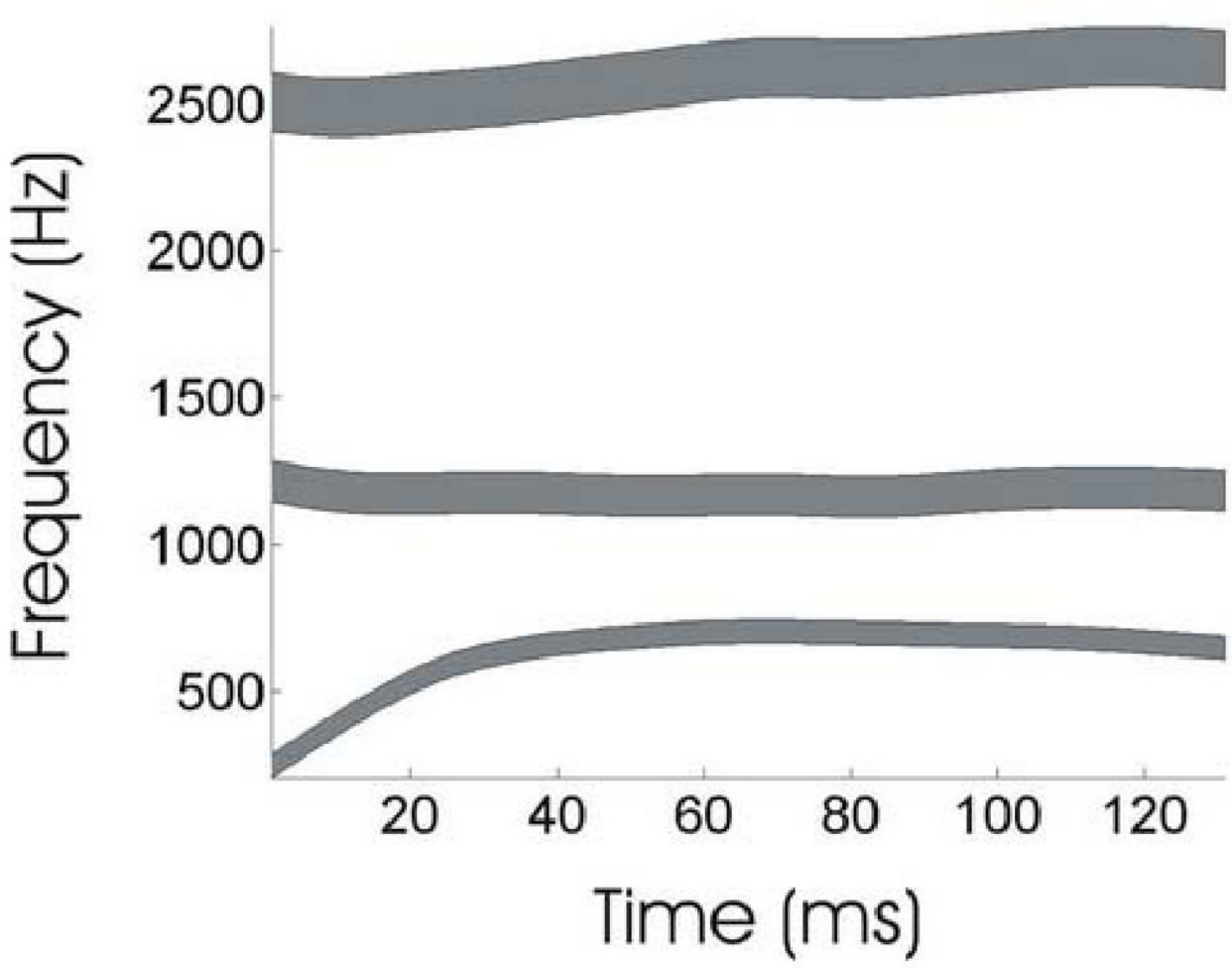

Figure 2.
Auditory target region for the first three formants of the syllable “ba” as learned by the model from an audio sample of an adult male speaker.

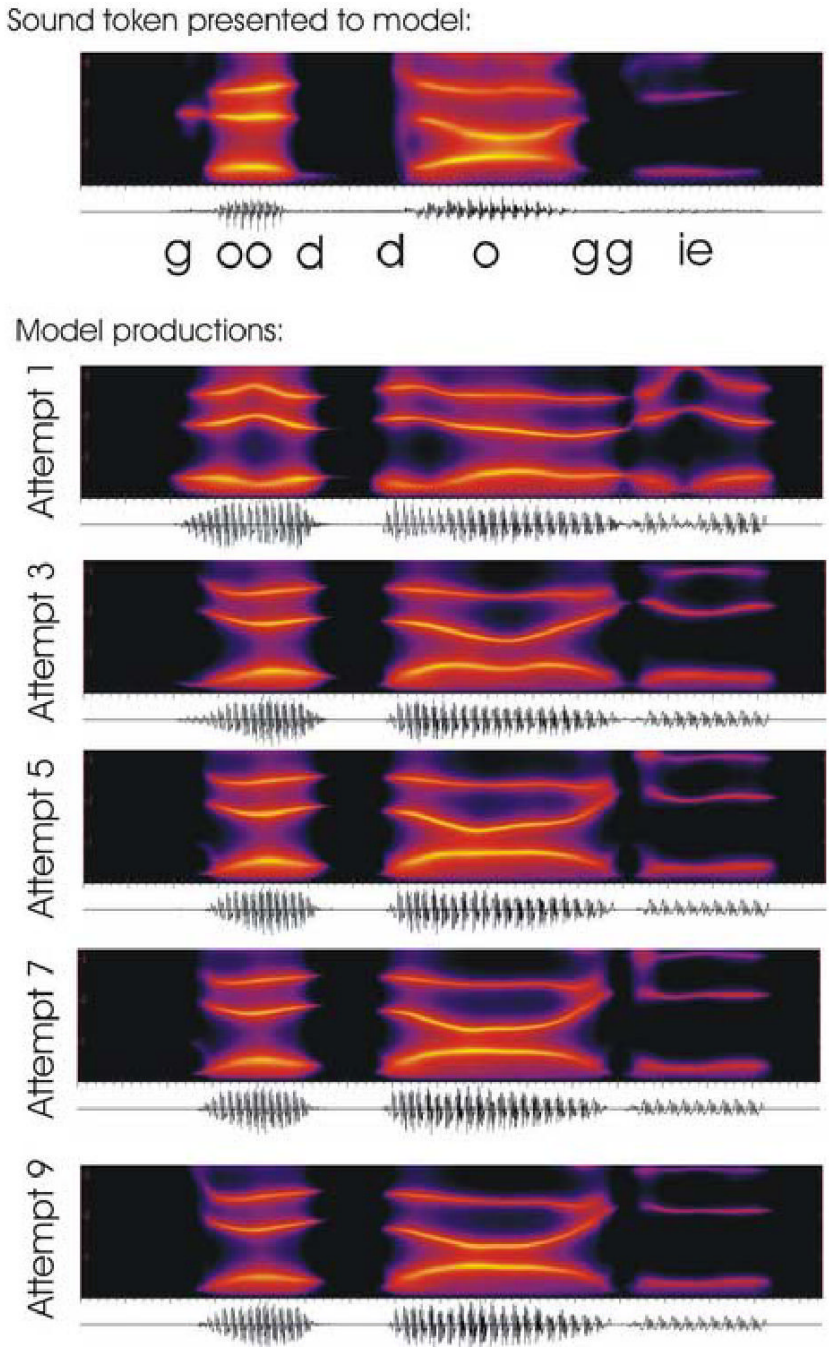

Figure 3.

Spectrograms showing the first three formants of the utterance “good doggie” as produced by an adult male speaker (top panel) and by the model (bottom panels). The model first learns an acoustic target for the utterance based on the sample it is presented (top panel). Then the model attempts to produce the sound, at first primarily under feedback control (Attempt 1), then with progressively improved feedforward commands supplementing the feedback control (Attempts 3, 5, 7, and 9). By the 9th attempt the feedforward control signals are accurate enough for the model to closely imitate the formant trajectories from the sample utterance.

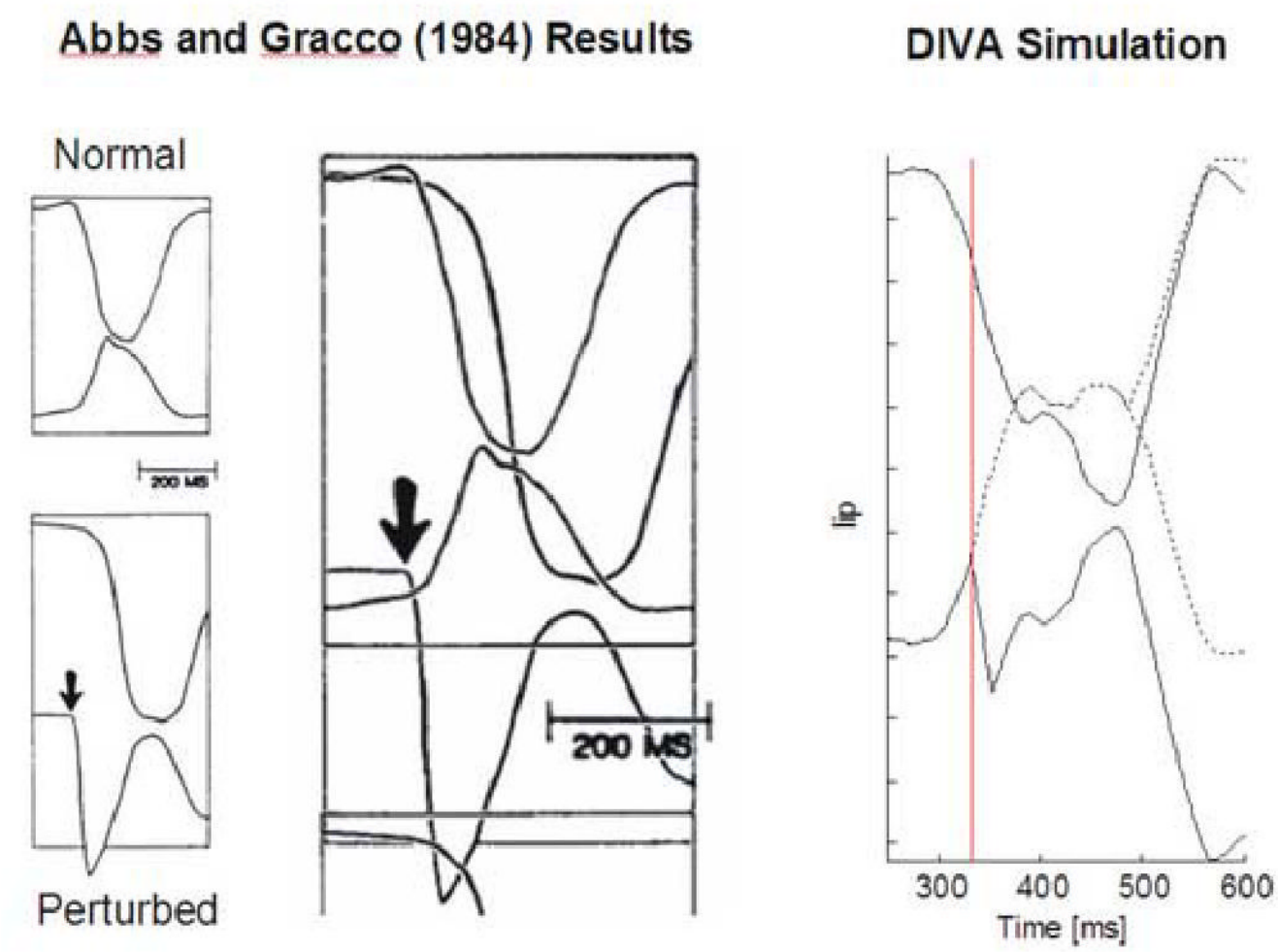

Figure 4.

Abbs and Gracco (1984) lip perturbation experimental results (left) and model simulation results (right). Far left panels show upper and lower lip positions during bilabial consonant production in the normal (top) and perturbed (bottom) conditions of the Abbs and Gracco (1984) experiment; shown to the right of this is a superposition of the normal and perturbed trials in a single image. Arrows indicate onset of perturbation. [Adapted from Abbs and Gracco (1984).] The right panel shows the lip heights from model simulations of the control (dashed lines) and perturbed (solid lines) conditions for the same perturbation, applied as the model starts to produce the /b/ in /aba/ (vertical line). The solid lines demonstrate the compensation provided by the upper and lower lips, which achieve contact despite the perturbation. The latency of the model's compensatory response is within the range measured by Abbs and Gracco (1984).

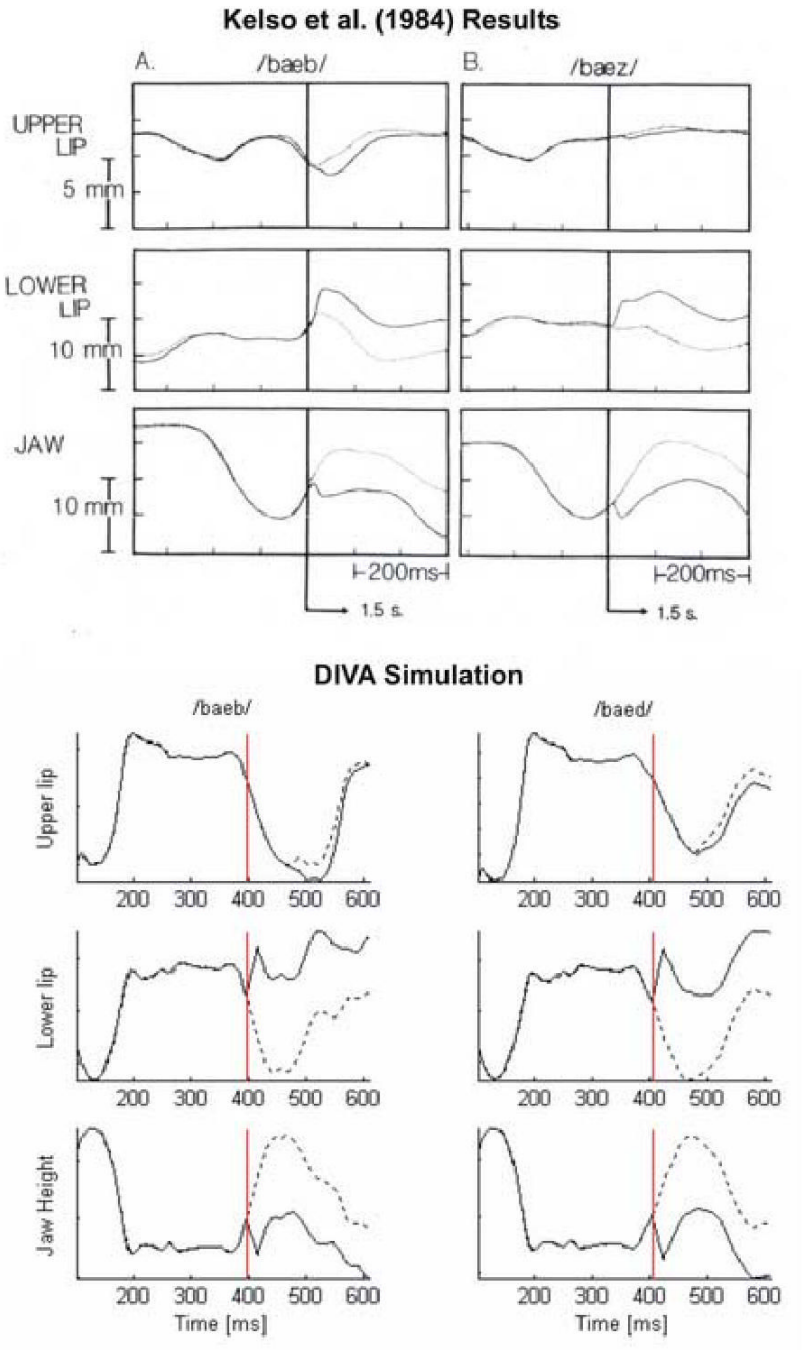

Figure 5.

Top: Results of Kelso et al. (1984) jaw perturbation experiment. Dotted lines indicate normal (unperturbed) trials, and solid lines indicate perturbed trials. The vertical line indicates onset of perturbation. Lower lip position is measured relative to jaw. Subjects produce compensatory downward movement of the upper lip for the bilabial stop /b/ but not for the alveolar stop /d/. [Adapted from (Kelso, Tuller, Vatikiotis-Bateson, & Fowler, 1984).] **Bottom:** Corresponding DIVA simulation. As in the Kelso et al. (1984) experiment, the model produces a compensatory downward movement of the upper lip for the bilabial stop /b/ but not for the alveolar stop /d/.

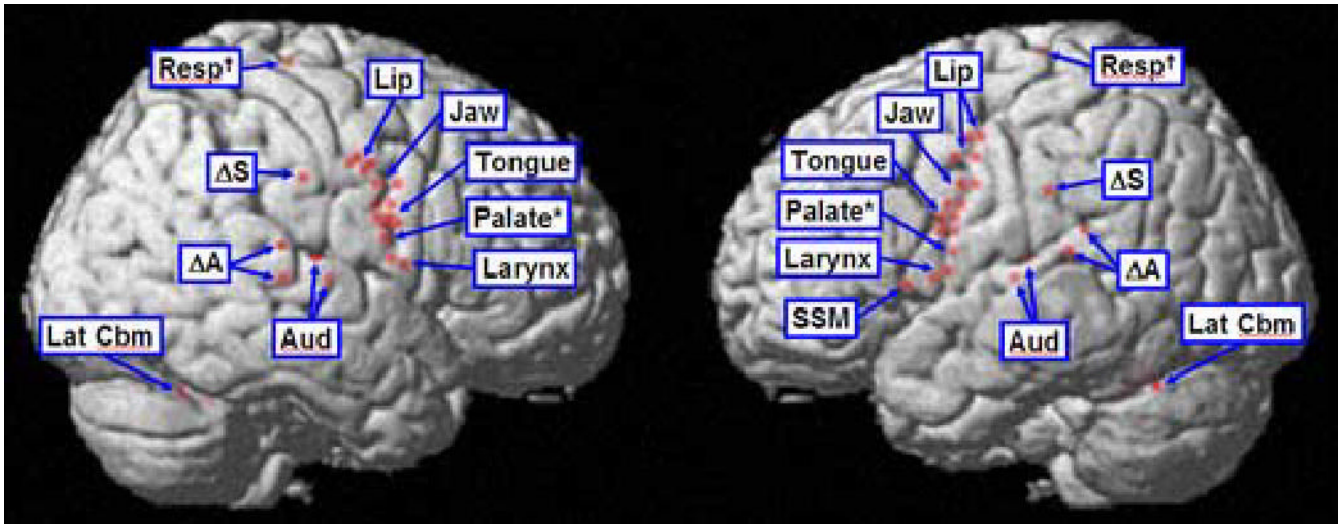

Figure 6.

Rendered lateral surfaces of the SPM standard brain indicating locations of the model components as described in the text. Medial regions (anterior paravermal cerebellum and deep cerebellar nuclei) are omitted. Unless otherwise noted, labels along the central sulcus correspond to a motor (anterior) and a somatosensory (posterior) representation for each articulator. Abbreviation key: Aud = auditory state cells; ΔA = auditory error cells; ΔS = somatosensory error cells; Lat Cbm = superior lateral cerebellum; Resp = motor respiratory region; SSM = speech sound map. *Palate representation is somatosensory only. †Respiratory representation is motor only.

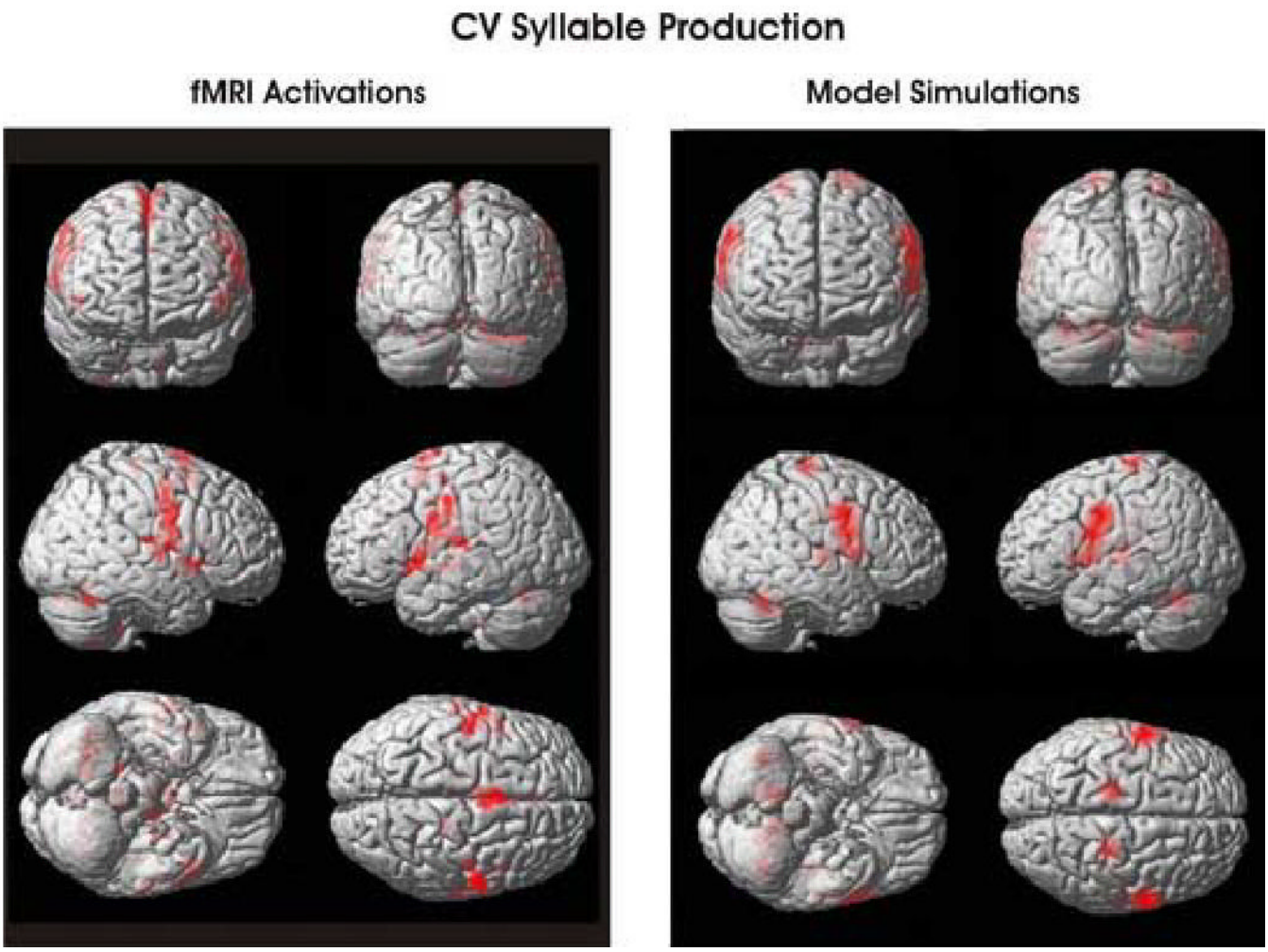

Figure 7.
fMRI activations measured in human subjects while they read simple syllables from a screen (left) and simulated fMRI activations derived from the model's cell activities during simple syllable production (right). See text for details.

Table 1

Montreal Neurological Institute (MNI) normalized spatial coordinates of DIVA model components mapped onto the left and right hemisphere of the canonical single brain provided with the SPM2 analysis software package. SPT = Sylvian-parietal-temporal region as described by Hickok et al. (2004).

Model Components	Left			Right		
	x	y	z	x	y	z
Motor Tongue						
1	−60.2	2.1	27.5	62.9	2.5	28.9
2	−60.2	3.0	23.3	66.7	2.5	24.9
3	−60.2	4.4	19.4	64.2	3	22
Motor Lip						
Upper	−53.9	−3.6	47.2	59.6	−7.2	42.5
Lower	−56.4	0.5	42.3	59.6	−3.6	40.6
Motor Jaw	−59.6	−1.3	33.2	62.1	3.9	34.0
Motor Larynx	−58.1	6.0	6.4	65.4	5.2	10.4
Motor Respiration	−17.4	−26.9	73.4	23.8	−28.5	70.1
Cerebellum						
Anterior Paravermis	−18	−59	−22	16	−59	−23
Anterior Lateral	−36	−59	−27	40	−60	−28
Deep Cerebellar Nuclei	−10.3	−52.9	−28.5	14.4	−52.9	−29.3
Speech Sound Map						
Inf. Prefrontal Gyrus	−56.5	14.8	4.8			
Sensory Tongue						
1	−60.2	−2.8	27.0	62.9	−1.5	28.9
2	−60.2	−0.5	23.3	66.7	−1.9	24.9
3	−60.2	0.6	20.8	64.2	0.1	21.7
Sensory Lip						
Upper	−53.9	−7.7	47.2	59.6	−10.2	40.6
Lower	−56.4	−5.3	42.1	59.6	−6.9	38.2
Sensory Jaw	−59.6	−5.3	33.4	62.1	−1.5	34.0
Sensory Larynx	−61.8	1	7.5	65.4	1.2	12
Sensory Palate	−58	−0.7	14.3	65.4	−0.4	21.6
Somatosensory Error Cells						
Supramarginal Gyrus	−62.1	−28.4	32.6	66.1	−24.4	35.2
Auditory State Cells						
Heschl's gyrus	−37.4	−22.5	11.8	39.1	−20.9	11.8
Planum temporale	−57.2	−18.4	6.9	59.6	−15.1	6.9
Auditory Error Cells						
SPT	−39.1	−33.2	14.3	44	−30.7	15.1
Post. Sup. Temporal Gyrus	−64.6	−33.2	13.5	69.5	−30.7	5.2