ASHA 2007 Zemlin Memorial Award Lecture: The Neural Control of Speech

Frank H. Guenther

Department of Cognitive and Neural Systems, Boston University, Boston, MA
and
Division of Health Sciences and Technology, Harvard University - Massachusetts Institute of Technology, Cambridge, MA

Abstract

Speech production involves coordinated processing in many regions of the brain. To better understand these processes, our research team has designed, tested, and refined a neural network model whose components correspond to brain regions involved in speech. Babbling and imitation phases are used to train neural mappings between phonological, articulatory, auditory, and somatosensory representations. After learning, the model can produce combinations of the sounds it has learned by commanding movements of an articulatory synthesizer. Computer simulations of the model account for a wide range of experimental findings, including data on acquisition of speaking skills, articulatory kinematics, and brain activity during speech. The model is also being used to investigate speech motor disorders, such as stuttering, apraxia of speech, and ataxic dysarthria. These projects compare the effects of damage to particular regions of the model to the kinematics, acoustics, or brain activation patterns of speakers with similar damage. Finally, insights from the model are being used to guide the design of a brain-computer interface for providing prosthetic speech to profoundly paralyzed individuals.

Introduction

The production of speech requires integration of diverse information sources in order to generate the intricate pattern of muscle activations required for fluency. Accordingly, a large portion of the cerebral cortex is involved in even the simplest speech tasks, such as reading a single syllable. There are three main types of information involved in the production of speech sounds: auditory, somatosensory, and motor, represented in the temporal, parietal, and frontal lobes of the cerebral cortex, respectively. These regions and their interconnections, along with subcortical structures such as the cerebellum, basal ganglia, thalamus, and brain stem, constitute the neural control system responsible for speech production.

Since 1992, our laboratory has developed and refined a neural network model that provides a quantitative account of the interactions between motor, somatosensory, and auditory cortical areas that underlie speech motor control. Because a central aspect of the model concerns how the brain transforms desired movement **D**irections in sensory space **I**nto **V**elocities of the **A**rticulators, the model is called the DIVA model. The model is implemented in computer simulations that control an articulatory synthesizer in order to produce an acoustic signal. The model's articulator movements and acoustic signal can be compared to the productions of human speakers to verify its ability to account for key aspects of human speech, and activity in the model's neurons can be compared to the results of neuroimaging studies of speech.

The following section provides an overview of the DIVA model, including a treatment of interactions between feedforward control mechanisms (or *motor programs*) and sensory feedback-based control mechanisms during speech production. Then, a brief treatment of how the model can be used to investigate disorders of speech motor control is provided, followed by a description of the design of a neural prosthetic device for speech production using insights from the model.

Overview of the DIVA Model

Figure 1 provides a schematic overview of the cortical interactions underlying speech production according to the DIVA model. Subcortical components of the model are omitted from the figure for clarity. Each box corresponds to a map of neurons located in a particular part of the brain. The pattern of activity in one of these maps at a given point in time provides a *neural representation* of information within a particular reference frame. For example, the *auditory state map* represents information about the acoustic signal arriving via the auditory periphery. The arrows between boxes represent transformations of information from one form of representation into another. These transformations are performed by synapses in the neural network that must be tuned through behavioral experience.

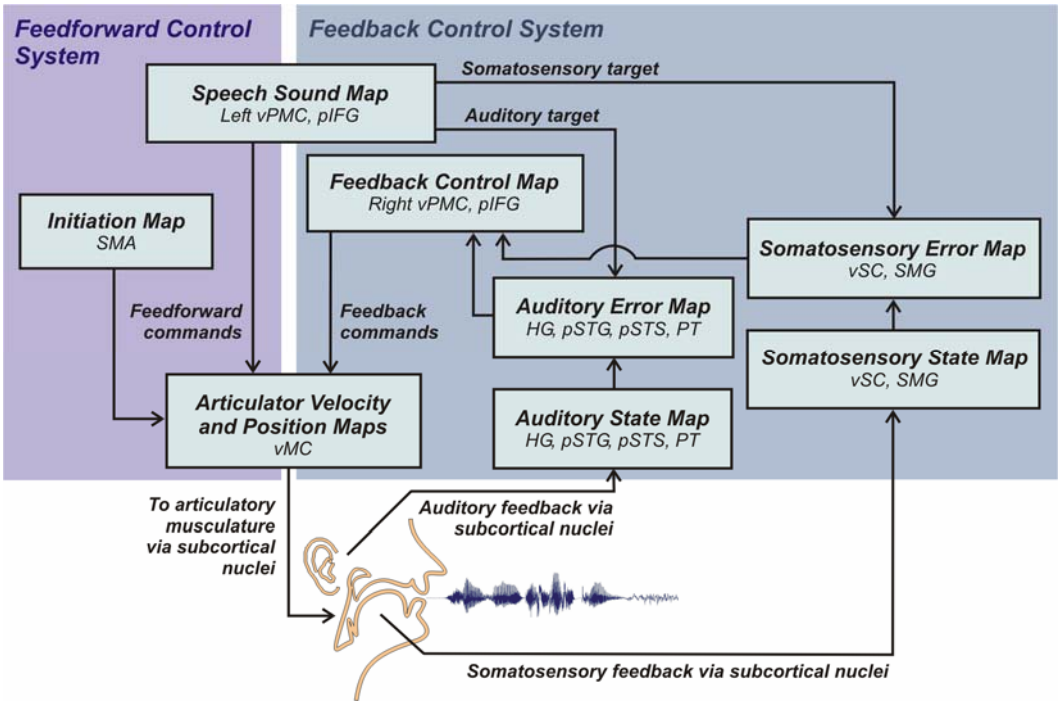

Figure 1. *Schematic of the cortical components of the DIVA model. [Abbreviations: HG=Heschl's gyrus; pIFG=posterior inferior frontal gyrus; pSTG=posterior superior temporal gyrus; pSTS=posterior superior temporal sulcus; PT=planum temporale; SMA=supplementary motor area; SMG=supramarginal gyrus; vMC=ventral motor cortex; vPMC=ventral premotor cortex; vSC=ventral somatosensory cortex.]*

According to the model, production of a speech sound starts with activation of cells in a *speech sound map* hypothesized to lie in the left ventral premotor cortex and posterior portion of the inferior frontal gyrus (Broca's area), combined with an initiation signal arriving from the supplementary motor area (SMA). A "speech sound" can be a phoneme, syllable, or whole word. Activation of the speech sound map cell leads to motor commands that arrive in ventral motor cortex via two control subsystems: a feedforward control subsystem (violet shading) and a feedback control subsystem (blue shading). The feedback control subsystem can be further broken into two components: an auditory feedback control subsystem and a somatosensory feedback control subsystem. The readout of speech motor commands from motor cortex to the motor periphery is gated on or off by an initiation signal arriving from the SMA. In other words, the ventral premotor cortex determines *what* motor commands to read out for a speech sound, while the SMA determines *when* to read them out.

Before it can produce speech sounds, the model must undergo a training process analogous to infant babbling and early word imitation. Synaptic projections between auditory, somatosensory, and motor cortical regions are tuned during a babbling phase in which semi-random articulator movements are used to produce auditory and somatosensory feedback. Learning in this stage is not specific to

particular speech sounds; the learned sensory-motor transformations are used for all speech sounds learned later.

In the next learning stage, the model is presented with sample speech sounds to learn, much like an infant is exposed to the sounds of his/her native language. These sounds take the form of time-varying acoustic signals corresponding to phonemes, syllables, or words spoken by a human speaker. Based on these samples, the model learns an auditory target for each sound, encoded in the synaptic projections from the speech sound map to the higher-order auditory cortical areas in the superior temporal lobe. The targets consist of time-varying regions that encode the allowable variability of the acoustic signal. The use of target regions (rather than point targets) is an important aspect of the model that provides a unified explanation for a wide range of speech production phenomena, including contextual variability, anticipatory coarticulation, carryover coarticulation, and speaking rate effects (Guenther, 1995).

Learning of a sound's auditory target involves activation of a speech sound map cell that will represent the sound for production. This in turn leads to tuning of the synapses projecting from that speech sound map cell to the auditory cortical areas. Later, the same speech sound map cell can be activated to generate the motor commands necessary for producing the sound. Thus, the speech sound map cells are activated both when perceiving a sound and, when producing the same sound. Neurons with this property, called *mirror neurons* (e.g., Rizzolatti, Fadiga, Gallese, & Fogassi, 1996), have been identified in monkey premotor area F5, which is believed to be the monkey homologue of Broca's area.

After an auditory target for a sound has been learned, the model can attempt to produce the sound. On the first attempt, the model will not have a tuned feedforward command for the sound; thus, it must rely heavily on the auditory feedback control subsystem for production. On each attempt, the feedforward command is updated to incorporate the commands generated by the auditory feedback control subsystem on that attempt. This results in a more accurate feedforward command for the next attempt. Eventually, the feedforward command becomes sufficient to produce the sound in normal circumstances. That is, the feedforward command is accurate enough that it generates no auditory errors during production of the sound and, thus, does not invoke the auditory feedback control subsystem. At this point, the model can fluently produce the speech sound.

As the model repeatedly produces a speech sound, it also learns a somatosensory target region for the sound, analogous to the auditory target region mentioned above. This target represents the expected tactile and proprioceptive sensations associated with the sound and is used in the somatosensory feedback control subsystem, where it is compared to incoming somatosensory information from the speech articulators to detect somatosensory errors. The corresponding *somatosensory state map* and *somatosensory error map* are hypothesized to reside in somatosensory cortical regions of the inferior parietal lobe, including the ventral

somatosensory cortex and anterior supramarginal gyrus. Unlike the auditory target for a sound, the somatosensory target cannot be learned in its entirety from watching and/or listening to another speaker, since crucial information such as tongue shape and tactile patterns are not directly available to the viewer/listener. Instead, the somatosensory target must be learned primarily by monitoring one's own correct self-productions, a learning process that occurs later than the learning of auditory targets in the model.

Computer simulations of the model have verified its ability to account for a wide range of acoustic and kinematic data concerning speech production (e.g., Guenther, 1995; Guenther, Hampson, & Johnson, 1998). Furthermore, because the model's components correspond to neurons and are given precise anatomical locations, activity in the model's cells can be compared to neuroimaging data (e.g., Guenther, Ghosh, & Tourville, 2006). The model thus provides a unified account of behavioral and neurophysiological aspects of speech production.

Investigating Communication Disorders

Each of the DIVA model's components is assigned to a particular region of the brain, including specification in a standardized stereotactic space (Guenther et al., 2006). The effects of damage in particular brain regions can, therefore, be simulated, and the acoustic and kinematic abnormalities of the model's productions can be compared to the productions of speakers with the same neurological impairment. This makes the model ideal for combined theoretical and experimental investigations of speech motor disorders and their treatment.

One such disorder is acquired apraxia of speech (AOS), which is most commonly associated with damage to the left inferior frontal gyrus (IFG) and/or surrounding cortical areas including the ventral premotor cortex. This region is proposed as the location of the speech sound map in the DIVA model (Figure 1). AOS is typically described as a disorder of speech motor programming, a characterization that fits well with the DIVA model's assertion that synaptic projections from speech sound map cells representing syllables in left IFG to primary motor cortex encode the feedforward commands, or motor programs, for these syllables. Therefore, damage to this region should result in damage to the motor programs for speech sounds. Notably, damage to the right hemisphere IFG does not typically impair the readout of speech motor programs. According to the DIVA model, right hemisphere IFG is important for feedback control (*feedback control map* in Figure 1), but not feedforward control (Tourville, Reilly, & Guenther, 2008). Once the model has learned to produce speech, damage to the feedback control system does not impair the readout of speech motor programs by the feedforward control system, thus accounting for the lack of apraxia with right hemisphere IFG damage. In ongoing collaborative work with Dr. Donald Robin and colleagues at the University of Texas Health Science Center in San Antonio, we are using the DIVA model in combination with neuroimaging to interpret

the effects of treatment programs for apraxia of speech in terms of changes in the underlying neural circuitry.

In a separate project with collaborators Dr. Ben Maassen and Hayo Terband of Radboud University in Nijmegen, The Netherlands, we have simulated the effects of childhood apraxia of speech (CAS) with the DIVA model. Like AOS, CAS is typically defined as a disorder of motor planning or programming, but the neurological impairment underlying CAS differs substantially from AOS. Whereas AOS patients usually have clear damage to the left inferior frontal gyrus and/or surrounding cortex, CAS patients typically do not have obvious structural damage to a particular portion of the brain. Instead, they suffer from more diffuse impairment of brain function. By modeling the effects of different forms of brain impairment (for example, localized damage to the left inferior frontal gyrus in AOS versus diffuse neurological impairment in CAS), more accurate descriptions of expected functional deficits in different subtypes of apraxia can be formulated and tested in experimental studies involving AOS and CAS patients.

Although not shown in Figure 1 for clarity, the DIVA model also posits functional roles for a number of subcortical structures involved in speech. For example, the model posits functional roles for several different sub-regions of the cerebellum, including a differentiation in function between medial and lateral regions of the superior cerebellar cortex. This type of detailed functional characterization may aid in forming a deeper understanding of ataxic dysarthria, a disorder arising from damage to the cerebellum (Spencer & Slocomb, 2007). This deeper understanding may in turn lead to improved treatment programs or prosthetic devices. An example of the latter is provided next.

Designing Neural Prosthetics for Speech

In collaboration with Dr. Philip Kennedy's research team at Neural Signals, Inc., we are developing a neural prosthesis for real-time speech synthesis using a brain-computer interface in a volunteer suffering from *locked-in syndrome*, which is characterized by complete paralysis of voluntary movement with intact cognition. A schematic of the system is provided in Figure 2. The system utilizes a permanently implanted recording electrode (Kennedy & Bakay, 1998) located in the volunteer's speech motor cortex. Action potentials from approximately 40-50 individual neurons are collected, preprocessed, and decoded into commands to a speech synthesizer that outputs sound to the volunteer via computer speakers.

The design of the system utilizes insights from the DIVA model. According to the model, the region of the brain implanted in the volunteer is responsible for transforming intended syllables, represented as auditory traces, into movement commands for the speech articulators. Accordingly, the neural decoder is designed to detect intended *formant frequencies*, which are key acoustic aspects of speech sounds that relate closely to the changing shape of the vocal tract. The decoded formant frequencies are then sent to a formant synthesizer to create an audible speech signal.

The entire process from action potentials to sound output takes less than 50 ms, which is approximately the delay from speech motor cortical activity to sound output in a neurologically normal speaker. The volunteer thus receives real-time feedback of the synthesizer "movements" commanded by his ongoing speech motor cortical activity.

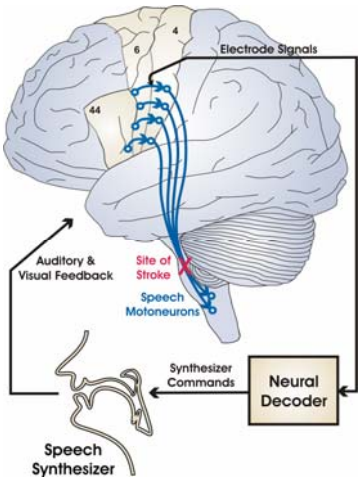

Figure 2. *Schematic of a brain-computer interface for real-time speech synthesis by a volunteer with locked-in syndrome due to brain stem stroke.*

The first real-time synthesis experiment with this system was performed in February 2008. On each trial of the experiment, the synthesizer was started at the formant frequencies for "uh," and the locked-in volunteer had to command changes in the formant frequencies to change the vowel to "ee," "ah," or "oo." The volunteer typically failed to produce a given vowel consistently in his first five attempts for that vowel, but by the sixth attempt his ability to control the synthesizer had improved to the point where he was successful on most subsequent attempts. In other words, the volunteer learned to control movements of the synthesizer to produce vowels (albeit crudely in this initial experiment), a process made possible by the availability of real-time audio feedback that allowed his speech motor system to tune up its commands to the synthesizer with practice. In ongoing research, we are pursuing ways to achieve faster learning and improved performance on a broader range of speech sounds, with the long-term goal of restoring conversational speech to profoundly paralyzed individuals via a portable, laptop-based decoding and synthesis system.

Summary

The DIVA model provides a detailed functional characterization of the neural processes underlying the production of speech sounds, including assignment of distinct functional roles to specific brain regions. The model serves as a theoretical framework for forming a detailed functional understanding of a number of communication disorders, and insights from the model are being used to help design and evaluate treatment programs and prosthetic devices for these disorders.

References

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

Guenther, F. H., Ghosh, S. S., & Tourville, J. A. (2006). Neural modeling and imaging of the cortical interactions underlying syllable production. *Brain and Language, 96,* 280-301.

Guenther, F. H., Hampson, M., & Johnson, D. (1998). A theoretical investigation of reference frames for the planning of speech movements. *Psychological Review, 105,* 611-633.

Kennedy, P. R., & Bakay, R. A. E. (1998). Restoration of neural output from a paralyzed patient by a direct brain connection. *NeuroReport,* 9, 1707-1711.

Rizzolatti, G., Fadiga, L., Gallese, V., & Fogassi, L. (1996). Premotor cortex and the recognition of motor actions. *Brain Research. Cognitive Brain Research, 3,* 131-141.

Spencer, K.A., & Slocomb, D.L. (2007). The neural basis of ataxic dysarthria. *The Cerebellum, 6,* 58-65.

Tourville, J. A., Reilly, K. J., & Guenther, F. H. (2008). Neural mechanisms underlying auditory feedback control of speech. *NeuroImage, 39,* 1429-1443.