THE NEURAL CONTROL OF SPEECH: FROM COMPUTATIONAL MODELING TO NEURAL PROSTHESIS

Frank H. Guenther

Departments of Speech, Language, & Hearing Sciences and
Biomedical Engineering, Boston University
guenther@bu.edu

ABSTRACT

Speech production is a complex sensorimotor task involving tightly coordinated processing in the frontal, temporal, and parietal lobes of the cerebral cortex. To better understand these processes, our laboratory has designed, experimentally tested, and iteratively refined a neural network model whose components correspond to the brain regions involved in speech. The model's components correspond to neural populations and are given precise anatomical locations. This allows activity in the model's neurons to be compared directly to neuroimaging data. 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 normal and perturbed speech. The model is also being used to investigate communication disorders, including stuttering, apraxia of speech, and spasmodic dysphonia. The model has also been used to guide development of a brain-computer interface aimed at restoring speech output to an individual suffering from locked-in syndrome.

Keywords: fMRI, neuroimaging, brain-computer interface, DIVA model.

1. INTRODUCTION

Since 1992, our laboratory has created, tested, and refined several neural network models of the brain mechanisms underlying the speech capacity in humans. One such model, the *D*irections *I*nto *V*elocities of *A*rticulators (DIVA) model, has become the most widely used account of the neural mechanisms responsible for speech motor control. The following sections introduce the DIVA model from a control systems perspective, describe the brain circuits that implement this speech motor controller, and provide examples of how the model serves as a guiding framework for investigating and treating communication disorders.

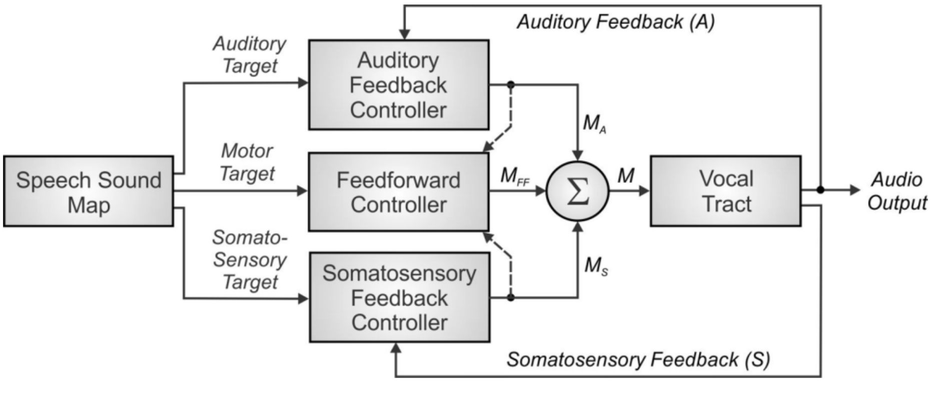
2. THE SPEECH CONTROL SYSTEM

Figure 1 illustrates the control system implemented by the DIVA model (e.g., [7], [4]). Production of a speech sound (which can be a word, syllable, or phoneme) begins with activation of a *speech sound map* node corresponding to that sound. This leads to the readout of three multidimensional signals representing the (i) *auditory target*, (ii) *somatosensory target*, and (iii) *motor target* for the chosen sound. Each signal is a function of time, covering the duration of the speech sound. These signals act as inputs to three semi-independent controllers: the *auditory feedback controller*, *somatosensory feedback controller*, and *feedforward controller*.

The auditory feedback controller compares the auditory target to the current auditory state. If the auditory state lies outside the auditory target region, auditory error signals are generated and translated into corrective motor commands (this process is not shown in Figure 1). These commands take the form of a vector of motor velocities, denoted M_A.

Similarly, the somatosensory feedback controller compares the somatosensory target of the sound to

Figure 1: The control scheme embodied by the DIVA model.



the current somatosensory state, generating somatosensory error signals if the somatosensory state is outside of the target region. These error signals are then mapped into corrective motor velocities, M_S.

The feedforward controller compares the motor target to the current motor state (not shown) and generates a motor velocity M_{FF}. Any corrective motor commands generated by the auditory and somatosensory feedback controllers are used to update the motor target for the next production attempt (indicated by dashed arrows).

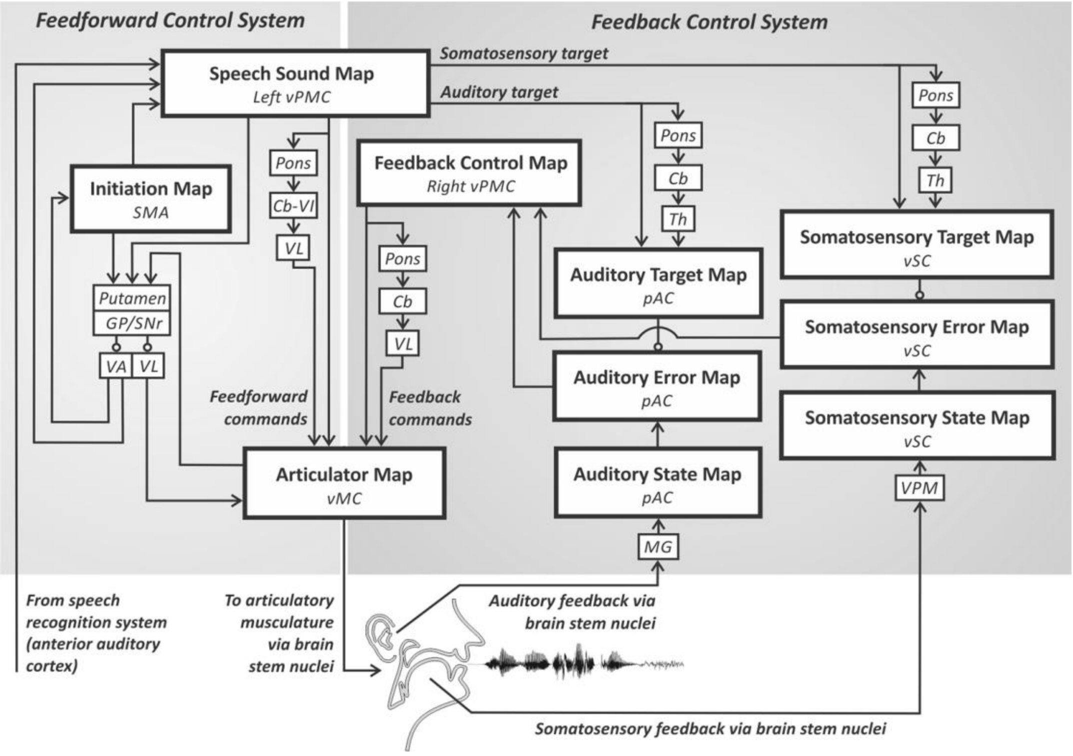
The outputs of these controllers are summed together to produce an overall motor command to the vocal tract, labelled M in Figure 1.

3. THE DIVA MODEL

The DIVA model implements the controller schematized in Figure 1 as an artificial neural network whose components correspond to specific regions of the brain. A schematic of the DIVA model is provided in Figure 2. Broadly speaking, the model's components are broken into a *feedforward control system* and a *feedback control system*, indicated by shading. Each box in Figure 2 corresponds to a set of neurons (or neural *map*) that represents a particular type of information in the model. Larger boxes represent cortical regions, and smaller boxes represent subcortical regions. Arrows correspond to excitatory axonal projections between neural maps, and lines terminating in circles represent inhibitory projections. Each model component (i.e., each box in Figure 2) is localized in the Montreal Neurological Institute (MNI) stereotactic reference frame, allowing the generation of fMRI-like brain activity from model simulations. The articulator movements, acoustic output, and brain activity produced by simulations of the model can be directly compared to the results of speech experiments.

The feedforward control system is responsible for generating previously learned motor programs for speech sounds. This process involves two components. The first component is responsible for launching the motor program at the appropriate instant in time. This is carried out by a cortico-basal ganglia-thalamo-cortical loop (*BG loop* hereafter) that involves an *initiation map* in the supplementary motor area (SMA) located on the medial wall of the frontal cortex. This loop is responsible for identifying the proper cognitive and sensorimotor context for producing the speech sound. For example, when saying the word "enter", the proper

Figure 2: Neural circuitry involved in speech motor control according to the DIVA model. [Abbreviations: Cb=cerebellum; Cb-VI=cerebellum lobule VI; GP=globus pallidus; MG=medial geniculate nucleus of the thalamus; pAC=posterior auditory cortex; SMA=supplementary motor area; SNr=substantia nigra pars reticula; Th=thalamus; VA=ventral anterior nucleus of the thalamus; VL=ventral lateral nucleus of the thalamus; vMC=ventral motor cortex; VPM=ventral posterior medial nucleus of the thalamus; vPMC=ventral premotor cortex; vSC=ventral somatosensory cortex.]



context for initiating the motor program for "ter" consists of (i) a cognitive context involving the desire to say the word "enter" and (ii) a sensorimotor context signaling the impending completion of articulation for "en". These contextual cues are monitored by the basal ganglia, with sensorimotor signals monitored by the putamen and cognitive signals monitored by the caudate nucleus (the latter not shown in Figure 2 as this component is outside the scope of the DIVA model). When the appropriate context for a sound is identified, a corresponding node is activated in the initiation map via the BG loop. Activation of the initiation map node initiates the readout of the learned motor program for the current speech sound.

The second component of the feedforward control system comprises the motor programs themselves, which are responsible for generating feedforward commands for producing learned speech sounds. These commands are encoded by synaptic projections from the *speech sound map* in left ventral premotor cortex (vPMC) to an *articulator map* in the ventral primary motor cortex (vMC) of the precentral gyrus bilaterally. The cortico-cortical projections from left vPMC to vMC are supplemented by a cerebellar loop passing through the pons, cerebellar cortex lobule VI (Cb-VI), and the ventral lateral (VL) nucleus of thalamus.

The feedback control system is broken into auditory and somatosensory subsystems. The auditory feedback control subsystem is responsible for detecting and correcting differences between the desired auditory signal for a speech sound and the current auditory feedback. According to the DIVA model, axonal projections from speech sound map nodes - both via cortical pathways and via a cerebellar loop involving the pons, cerebellum (Cb), and thalamus (Th) - to an *auditory target map* in the higher-order auditory cortical areas in posterior auditory cortex (pAC), including the planum temporale and posterior superior temporal gyrus and sulcus. These projections encode the expected auditory signal for the speech sound currently being produced. Activity in the auditory target map thus represents the auditory feedback that should arise when the speaker hears himself/herself producing the current sound. The targets consist of time-varying regions (or ranges) that encode the allowable variability of the acoustic signal throughout the syllable. The use of target *regions* (rather than point targets) is an important aspect of the DIVA model which 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 [5].

The auditory target for the current sound is compared to incoming auditory information from the auditory periphery; this information projects to cortex via the medial geniculate nucleus (MG) of the thalamus and is represented in the model's *auditory state map*. If the current auditory feedback is outside the target region, auditory error nodes in the higher-order auditory cortical areas become active (*auditory error map* in Figure 2). Like the auditory target map, the auditory state and error maps are hypothesized to lie in pAC. Auditory error node activities are then transformed into corrective motor commands through projections from the auditory error nodes to the *feedback control map* in right vPMC, which in turn projects to the articulator map in vMC both directly and via a loop through the pons, Cb, and VL.

The main components of the somatosensory feedback control subsystem are hypothesized to reside in ventral somatosensory cortex (*vSC*), including the ventral postcentral gyrus and the supramarginal gyrus. Projections from the speech sound map to the *somatosensory target map* (including a cerebellar loop projection) encode the expected somatosensory feedback (i.e., tactile and proprioceptive feedback arising from mechanoreceptors and muscle spindles in the vocal tract) during sound production. The model's *somatosensory state map* represents tactile and proprioceptive information from the speech articulators, which arrives from cranial nerve nuclei in the brain stem via the ventral posterior medial nucleus of the thalamus. Nodes in the *somatosensory error map* become active during speech if the speaker's somatosensory state deviates from the somatosensory target region for the sound being produced. The output of the somatosensory error map then propagates to the feedback control map to transform somatosensory errors into motor commands that correct those errors.

4. A FRAMEWORK FOR INVESTIGATING SPEECH DISORDERS

The DIVA model provides a powerful framework for studying disorders of speech communication that arise from neural causes. The following examples illustrate different ways in which the model can provide insights into these disorders.

One way the model contributes to studies of communication disorders is by providing mechanistic interpretations of neurological and/or neuroimaging findings. For example, Peeva et al. [11] studied the speech networks of high-functioning individuals with autism using diffusion tensor imaging and probabilistic tractography. The investigators noted anomalous axonal tracts between SMA and vPMC in the left hemisphere of autistic individuals compared to neurotypical controls. According to the DIVA model, this pathway is involved in initiating the readout of speech motor programs, as described above. This leads to the prediction that this anomaly may be related to reduced speech output in autism, a prediction that is being investigated in ongoing neuroimaging studies involving minimally verbal, language impaired, and high-functioning individuals with autism.

A second way the model can be used to investigate communication disorders is through simulations of "damaged" versions of the model that aim to mimic the effects of a neurological disorder. For example, Civier et al. [2], [3] used the DIVA model to investigate several hypotheses concerning possible neural deficits underlying stuttering. Simulations of the impaired versions of the model were then compared to behavioral and neuroimaging data to test between different possible accounts of the neural origins of stuttering. These simulations suggest that stuttering is likely not caused by an over-active auditory feedback control system, instead pointing to disordered functioning in the BG loop responsible for initiating speech motor programs (cf. [1]). Ongoing research is attempting to refine this account through improved modelling of basal ganglia circuitry and refined neuroimaging studies of individuals who stutter.

Finally, the model can be used as a guide for developing therapeutic interventions for speech disorders. In one example of this type, Guenther et al. [6] developed a brain-computer interface that allowed an individual with locked-in syndrome to produce vowels with a speech synthesizer. Neural signals from an electrode implanted in the individual's speech motor cortex were translated into formant frequency trajectories that were used to drive a formant synthesizer [10] in real time. The use of formant frequencies as the decoded variables was inspired by the DIVA model, in particular the fact that movement trajectories for speech sounds are planned within an auditory reference frame (e.g., formant space) rather than an articulatory reference frame. The model's prediction of a formant frequency representation in speech motor cortex was verified by neural recordings collected while the individual attempted to produce a series of vowel sounds (see Figure 4).

Figure 5 summarizes performance when the locked-in participant used the brain-computer interface to produce vowels in real time. Each session was divided into 4 blocks of approximately 5-10 vowel production attempts. Figure 5 denotes the mean production error measured in formant space across 25 sessions for each block. The individual's performance steadily improved over the course of an experimental session, with average formant error decreasing from approximately 420 Hz in the first block to approximately 230 Hz in the fourth block. This improvement in performance was made possible through real-time auditory feedback from the brain-computer interface during production attempts.

Figure 4: Preferred directions in formant frequency space (arrows) for neural units recorded from a locked-in individual attempting to produce a vowel sequence. Arrow length signifies the strength of the unit's directional preference (see [6]for details).

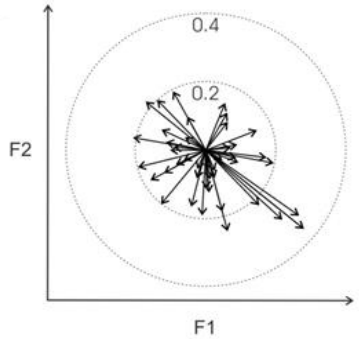
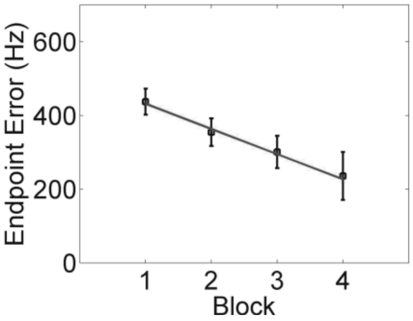


Figure 5: Performance of a real-time speech synthesis brain-computer interface for vowel production by a locked-in participant (see [6]).



6. CONCLUDING REMARKS

The proliferation of neuroimaging data in the past 25 years has provided a wealth of information regarding the neural mechanisms of speech in neurotypical individuals, as well as neural anomalies that accompany speech disorders. The DIVA model provides a neural and computational framework for interpreting these findings. Simulations of damaged versions of the model can be used to generate predictions that can be tested with behavioral and neuroimaging experiments, allowing quantitative testing between competing accounts of a disorder. Such simulations also provide a means for distinguishing primary deficits from secondary or compensatory effects. In recent years, additional computational models of the neural basis of speech production have arisen in the literature (e.g., [9], [8]). Future studies that test between these accounts will lead the field toward an ever-refined understanding of the neural underpinnings of speech and its disorders. This in turn will foster the development of treatments that either normalize neural function or make optimal use of unimpaired neural circuitry.

7. ACKNOWLEDGMENT

This research supported by grants R01 DC002852 and R01 DC007683 from the National Institute on Deafness and Other Communication Disorders and by the Center of Excellence for Learning in Education, Science, and Technology (CELEST), a National Science Foundation Science of Learning Center (NSF SMA-0835976).

8. REFERENCES

[1] Alm, P. A. 2004. Stuttering and the basal ganglia circuits: a critical review of possible relations. *J. Commun. Disord.* 37, 325-369.

[2] Civier, O., Bullock, D., Max, L., Guenther, F. H. 2013. Computational modeling of stuttering caused by impairments in a basal ganglia thalamo-cortical circuit involved in syllable selection and initiation. *Brain Lang.* 126, 263-278.

[3] Civier, O., Tasko, S. M., Guenther, F. H. 2010. Overreliance on auditory feedback may lead to sound/syllable repetitions: Simulations of stuttering and fluency-inducing conditions with a neural model of speech production. *J. Fluency Disord.* 35, 246-279

[4] Golfinopoulos, E., Tourville, J. A., Guenther, F. H. 2010. The integration of large-scale neural network modeling and functional brain imaging in speech motor control. *NeuroImage* 52, 862-874.

[5] Guenther, F. H. 1995. Speech sound acquisition, coarticulation, and rate effects in a neural network model of speech production. *Psychol. Rev.* 102, 594-621.

[6] Guenther, F. H., Brumberg, J. S., Wright, E. J., Nieto-Castanon, A., Tourville, J. A., Panko, M., Law, R., Siebert, S. A., Bartels, J. L., Andreasen, D. S., Ehirim, P., Mao, H., Kennedy, P. R. 2009. A wireless brain-machine interface for real-time speech synthesis. *PLoS ONE* 4, e8218+.

[7] Guenther, F. H., Ghosh, S. S., Tourville, J. A. 2006. Neural modeling and imaging of the cortical interactions underlying syllable production. *Brain Lang.* 96, 280-301.

[8] Hickok, G. 2012. Computational neuroanatomy of speech production. *Nat. Rev. Neurosci*. 13, 135-145.

[9] Houde, J. F. Nagarajan, S. S. 2011. Speech production as state feedback control. *Front. Hum. Neurosci.* 5, 82.

[10] Klatt D. H. 1980. Software for a cascade/parallel formant synthesizer. *J. Acoust. Soc. Am.* 67, 971–995.

[11] Peeva, M. G., Tourville, J. A., Agam, Y., Holland, B., Manoach, D.S., Guenther, F.H. 2013. White matter impairment in the speech network of individuals with autism spectrum disorder. *NeuroImage Clin.* 3, 234-241.