A NEURAL MODEL OF SPEECH PRODUCTION

Frank H. Guenther[1,2], Satrajit S. Ghosh[1], & Alfonso Nieto-Castanon[1]

[1] Department of Cognitive & Neural Systems, Boston University
[2] Research Laboratory of Electronics, Massachusetts Institute of Technology

ABSTRACT: This paper describes the most recent version of the DIVA model, a neural network model of the brain computations underlying the acquisition and production of speech sounds. The model, which is implemented as a set of equations representing neural activity and synaptic strengths, is designed to account for the results of functional magnetic resonance imaging (fMRI) and electromagnetic midsagittal articulometry (EMMA) experiments concerning the production of speech. Computer simulations of the model have been performed to illustrate its ability to account for speaker-specific articulator movements in different phonetic contexts, as well as fMRI activations seen during normal and perturbed speech. The model is also used to generate predictions that guide new fMRI and EMMA experiments aimed at achieving a better understanding of the neural bases of speech. The results of these experiments are in turn used to further refine the model. Finally, the model can be used to investigate the effects of various types of neurological damage on speaking skills.

INTRODUCTION: OVERVIEW OF THE DIVA MODEL

Figure 1 provides an overview of the DIVA model (e.g., Guenther, 1994; Guenther et al., 1995; 1998; Perkell et al., 2000; Callan et al., 2000). The model consists of a neural network controller whose cells correspond to boxes and synaptic weights correspond to arrows in the figure. The neural network utilizes a babbling stage to learn the neural mappings (arrows) necessary for controlling an articulatory synthesizer (e.g., Maeda, 1990). The output of the model (labelled "To Muscles" in Figure 1) specifies the positions of the 7 articulators that determine the vocal tract shape in the articulatory synthesizer.

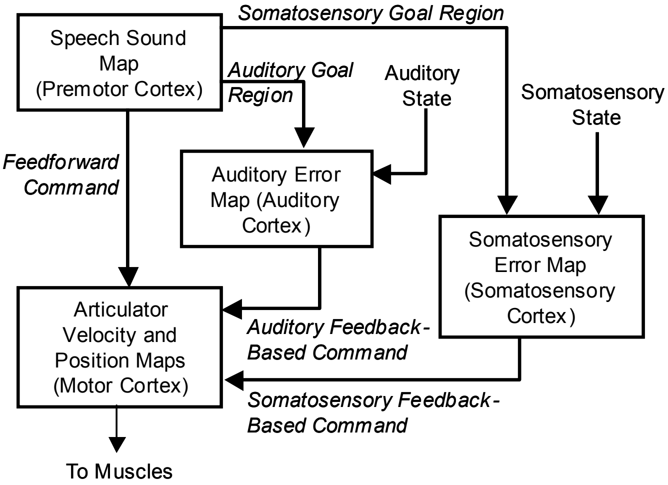
Somatosensory Goal Region
Speech Sound Map (Premotor Cortex)
Auditory Goal Region
Auditory State
Somatosensory State
Feedforward Command
Auditory Error Map (Auditory Cortex)
Somatosensory Error Map (Somatosensory Cortex)
Articulator Velocity and Position Maps (Motor Cortex)
Auditory Feedback-Based Command
Somatosensory Feedback-Based Command
To Muscles

Figure 1. Overview of the hypothesized neural processing stages involved in speech production according to the DIVA model.

In the model, production of a phoneme or syllable starts with activation of a speech sound map cell corresponding to the sound to be produced. These cells are hypothesized to correspond to "mirror neurons" that have been found in numerous studies of premotor cortex, including studies of speech. They can also be interpreted as part of a mental syllabary as described by Levelt and colleagues (e.g., Levelt and Wheeldon, 1994). After a speech sound map cell has been activated, signals from the premotor cortex travel to the auditory and somatosensory cortical areas through tuned synapses that encode sensory expectations for the sound being produced. These "forward models" are hypothesized to include both cortical and cerebellar components, with the cerebellar contribution being particularly important for fine temporal details. The synapses projecting from the premotor cortex to the higher-order 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 and/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 this discrepancy via projections from auditory to motor cortical areas. A second set of synapses, projecting from the premotor cortex to the higher-order somatosensory cortical areas, encodes 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 and averaging these somatosensory consequences 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-based 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 "inverse model" projections are tuned during babbling by monitoring the relationship between movement commands and their sensory consequences. 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.

MODEL EQUATIONS

In this section we present the equations that define cell activities in the model.

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

$$(1) \qquad Au(t) = f_{AcAu}(Acoust(t - \tau_{AcAu}))$$

where f_{AcAu} is the function that transforms an acoustic signal into the corresponding auditory 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. The model also has ***auditory error cells*** in these same cortical regions that encode the difference between auditory target regions for the sound being produced and the current auditory state as represented by *Au*(*t*). The activity of the auditory error cells (ΔAu) is defined by the following equation:

$$(2) \qquad \Delta Au(t) = Au(t) - P(t - \tau_{PAu})z_{PAu}(t) - M(t - \tau_{MAu})z_{MAu}(t)$$

where τ_{PAu} and τ_{MAu} are the propagation delays for the signals from premotor and motor cortex to auditory cortex, and z_{MAu} and $z_{PAu}(t)$ are synaptic weights[1] 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 sounds being produced.

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

[1] The synaptic weights in these equations are effective synaptic weights; that is, they depend on variables other than time (e.g., the current motor state) and don't represent single synapses in the neural network. A more thorough treatment of this issue is beyond the scope of the current paper.

$$(3) \qquad S(t) = f_{ArS}(Artic(t - \tau_{ArS}))$$

where f_{ArS} is a function that transforms the current state of the articulators into the corresponding somatosensory map state (e.g., positions and velocities of articulators).

Somatosensory error cells code the difference between the somatosensory target region for a speech sound and the current somatosensory state:

$$(4) \qquad \Delta S(t) = S(t) - P(t - \tau_{PS})z_{PS}(t) - M(t - \tau_{MS})z_{MS}(t)$$

where τ_{PS} and τ_{MS} are the propagation delays from premotor and motor cortex to somatosensory cortex, and $z_{MS}(t)$ and $z_{PS}(t)$ 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.

According to the model, feedforward and feedback-based control signals are combined in motor cortex. The model's ***motor cortex velocity cells*** correspond to "phasic" cells found in motor cortex single-cell recording studies. The model includes two sets of motor velocity cells: one that encodes a feedforward control signal and one that encodes a feedback control signal.

Feedback control signals project from sensory error cells to the motor cortex. These "inverse model" projections are governed by the following equation:

$$(5) \qquad \dot{M}_{Feedback}(t) = \Delta Au(t - \tau_{AuM})z_{AuM} + \Delta S(t - \tau_{SM})z_{SM}$$

where z_{AuM} and z_{SM} are synaptic weights that transform directional sensory error signals into motor velocities that correct for these errors. 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 the pseudoinverse of the Jacobian of the function relating articulator positions (*M*) to the corresponding sensory state (*Au, S*). These weights can be tuned during babbling by monitoring the relationship between movement commands and their sensory consequences.

The feedforward motor command, hypothesized to project from ventrolateral premotor cortex to primary motor cortex both directly and via the cerebellum, is represented by the following equation in the model:

$$(6) \qquad \dot{M}_{Feedforward}(t) = P(t - \tau_{PM})z_{PM}(t) - M(t)\ .$$

The weights $z_{PM}(t)$ encode the feedforward motor command for the speech sound being produced. This command can be learned over time by averaging the motor commands from previous attempts to produce the sound.

The model's ***motor cortex position cells*** correspond to "tonic" cells found in motor cortex single-cell recording studies. They represent the length of a muscle or muscle synergy, and they act as a command to the motor periphery. Their activity is governed by the following equation:

$$(7) \qquad M(t) = M(0) + \alpha_{ff}\int_0^t \dot{M}_{Feedforward}(t)g(t)dt + \alpha_{fb}\int_0^t \dot{M}_{Feedback}(t)g(t)dt$$

where α_{fb} and α_{ff} are parameters that determine how much the model is weighted toward feedback control and feedforward control, respectively, and $g(t)$ is a speaking rate signal that is 0 when not speaking and 1 when speaking at a maximum rate.

Before an infant has any practice producing a speech sound, the contribution of the feedforward control signal to the overall motor command should be small since it will not yet be tuned. Therefore, during the first few productions, the primary mode of control will be feedback-based 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 and 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 (see Callan et al., 2000).

The model also contains variables corresponding to the current articulatory and acoustic state. These values do not correspond to any brain cell activities; they correspond instead to the physical positions of the articulators and the resulting acoustic signal. The articulatory state describes the positions of the seven articulators in the Maeda articulatory synthesizer (Maeda, 1990), and is governed by the following equation in the model:

$$(8) \qquad Artic(t) = f_{MAr}(M(t-\tau_{MAr})) + Pert(t)$$

where f_{MAr} is the function relating the motor cortex position command to the Maeda parameter values, 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. The acoustic state is determined from the articulatory state as follows:

$$(9) \qquad Acoust(t) = f_{ArAc}(Artic(t))$$

where f_{ArAc} is the transformation performed by Maeda's articulatory synthesis software.

TESTING THE MODEL WITH ARTICULOMETRY

Numerous computer simulations have demonstrated that the model provides a straightforward, unified account for many aspects of speech movement kinematics (e.g., Guenther 1994; 1995; Guenther et al., 1998; 1999). In this section we describe one such study involving American English /r/ production.

The top row of Figure 2 shows the results of an electromagnetic midsagittal articulometry (EMMA) study of articulations for the phoneme /r/ in different phonetic contexts for one speaker (Guenther et al., 1999). The gesture for the /r/ in each context can be deduced by looking at the progression from the dashed outline (representing the tongue shape 75 ms before the acoustic center of the /r/) to the bold, solid outline (representing the tongue shape at the acoustic center of the /r/). Two things are of particular note: (1) the subject uses a wide range of tongue shapes for /r/ in different phonetic contexts, and (2) the gestures used to produce these tongue shapes differ greatly across contexts.

To directly test the control scheme represented by the DIVA model, we collected structural MRI scans of one of the speakers from the EMMA study while this speaker produced different phonemes. These scans were used to build an articulatory synthesizer matching the speaker's vocal tract shape, articulatory degrees of freedom, and acoustics (Nieto-Castanon et al., 2003). This speaker-specific vocal tract model was then controlled by the DIVA model after a babbling stage in which the model learned to control movements of the vocal tract. We then had the model produce the same phoneme strings as the speaker in the EMMA study and compared its movements to those of the modeled speaker. The results of these simulations are provided in the bottom row of Figure 2.

The model's gestures in each phonetic context closely mimic those of the speaker, as do the final tongue shapes for /r/. This result occurs in the model because the sound /r/ is represented primarily by an auditory target, without a corresponding vocal tract shape target. When the model moves toward this target from different initial configurations, different gestures and tongue shapes result for /r/, and these tongue shapes closely mimic those of the corresponding speaker. We take this to be strong evidence for the model's control scheme as a model of human movement control.

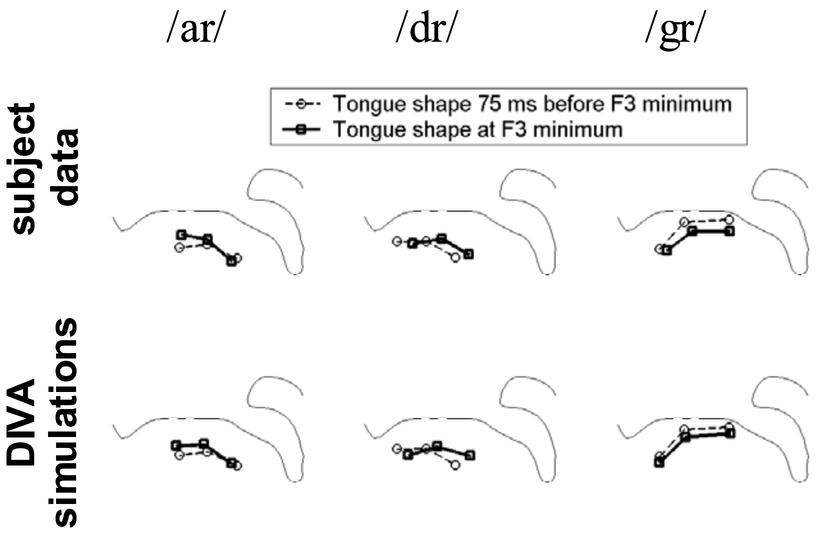

Figure 2. Tongue movements during /r/ production in an electromagnetic articulometry study (top) and in simulations of the DIVA model (bottom) in three different phonetic contexts: /ar/, /dr/, and /gr/. The lips are located to the left in each panel (velum outline is on right).

TESTING THE MODEL WITH FMRI

Because the model is specified as a neural network whose components relate closely to brain regions, it is possible to test predictions of the model using brain imaging techniques such as fMRI. Figure 3 summarizes the results of one such study, investigating the effects of unexpected jaw perturbation on brain activity during speech production.

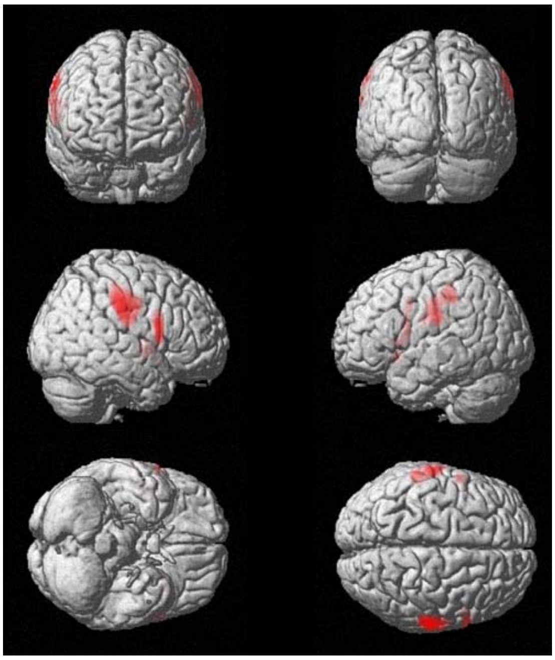

Figure 3. Brain activity (as measured with fMRI) resulting from unexpected jaw perturbation during speech production (average of 3 subjects).

The DIVA model predicts that jaw perturbation should cause increased activation in somatosensory cortex and supramarginal gyrus (due to somatosensory error signals) and motor cortex (due to feedback-based motor commands). These show significant activity in Figure 3, with a right hemisphere bias (see also Baciu et al., 2000). The model also predicts that activity may arise in auditory cortex if the jaw perturbation is disruptive enough to perturb the acoustic output of the vocal tract. A small amount of right hemisphere auditory cortical activation is evident in Figure 3, though the weakness of the activity and small number of subjects suggests caution in interpreting this activity.

CONCLUDING REMARKS

We are currently conducting experiments to test further predictions of the DIVA model with both articulometry at the MIT Research Laboratory of Electronics (in collaboration with Joseph Perkell) and fMRI at Massachusetts General Hospital. When necessitated by experimental findings, the model will be modified to better capture the available data. We believe this synthesis of modelling and experimentation will continue to produce insights into the neural bases of speech motor control.

ACKNOWLEDGEMENTS

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

REFERENCES

Baciu, M., Abry, C., & Segebarth, C. (2000) Equivalence motrice et dominance hémisphérique: Le cas de la voyelle [u]. Étude IRMf. *Actes des XXIIIèmes Journées d'Etude sur la Parole, Aussois, France*, pp. 213-216.

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

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

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., Espy-Wilson, C.Y., Boyce, S.E., Matthies, M.L., Zandipour, M., & Perkell, J.S. (1999) "Articulatory tradeoffs reduce acoustic variability during American English /r/ production" *Journal of the Acoustical Society of America* **105** 2854-2865.

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.

Kawato, M. & Gomi, H. (1992) "A computational model of four regions of the cerebellum based on feedback-error learning", *Biological Cybernetics* **68** 95-103.

Levelt, W.J. & Wheeldon, L. (1994) "Do speakers have access to a mental syllabary?" *Cognition* **50** 239-269.

Maeda, S. (1990) "Compensatory articulation during speech: evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model" in *Speech production and speech modelling* (W. J. Hardcastle and A. Marchal, eds.), p. 131-149. Boston: Kluwer Academic Publishers.

Nieto-Castanon, A., Guenther, F.H., Perkell, J.S., and Curtin, H.D. (2003) "A modeling investigation of articulatory variability and acoustic stability during American English /r/ production." Submitted for publication.

Perkell, J.S., Guenther, F.H., Lane, H., Matthies, M.L., Perrier, P., Vick, J., Wilhelms-Tricarico, R., & Zandipour, M. (2000) "A theory of speech motor control and supporting data from speakers with normal hearing and profound hearing loss" *Journal of Phonetics* **28** 233-272.