A NEURAL MODEL OF SPEECH PRODUCTION AND SUPPORTING EXPERIMENTS

Frank H. Guenther[1,2,3] & Joseph S. Perkell[2,1]

[1] Dept. of Cognitive & Neural Systems, Boston University, Boston, MA, USA
[2] Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, MA, USA
[3] Athinoula A. Martinos Center for Biomedical Imaging, Massachusetts General Hospital, Charlestown, MA, USA
guenther@cns.bu.edu

ABSTRACT

This paper describes the DIVA model of speech production and presents results of experiments designed to test and refine the model. According to the model, production of a phoneme or syllable starts with activation of a speech sound map cell (in left ventral premotor cortex) corresponding to the sound to be produced. This leads to production of the sound through two motor subsystems: a feedback control subsystem and a feedforward control subsystem. In the feedback control subsystem, 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 expectations take the form of time-varying auditory and somatosensory target regions. The target regions are compared to the current auditory and somatosensory state, and any discrepancy between the target and the current state leads to a corrective command signal to motor cortex. In the feedforward control subsystem, signals project from premotor cortex to primary motor cortex, both directly and via the cerebellum. These signals are tuned with practice by monitoring the commands from previous attempts to produce the sound, initially under feedback control. Feedforward and feedback-based control signals are combined in the model's motor cortex to form the overall motor command. We present experimental results that support two theoretical characteristics of the model: its use of auditory target regions (including hypothesized effects of perceptual acuity on production target size), and its ability to achieve stable acoustic results using motor equivalent tradeoffs between articulatory gestures.

INTRODUCTION: OVERVIEW OF THE DIVA MODEL

Figure 1 schematizes the DIVA model (e.g., Guenther, 1994, 1995; Guenther et al., 1998; Guenther and Ghosh, 2003), a neural network controller that utilizes a babbling stage to learn the sensorimotor transformations necessary for controlling a simulated vocal tract, or articulatory synthesizer (e.g., Maeda, 1990), in order to produce words, syllables, or phonemes. The output of the model specifies the positions of eight speech articulators that determine the vocal tract shape in the articulatory synthesizer. Each block in Figure 1 corresponds to a set of neurons that constitute a neural representation, or "map". Transformations between neural representations are carried out by filtering cell activations in one map through synapses projecting to another map; these synaptic projections are indicated by arrows in the figure. Model parameters, corresponding to 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 train mappings between different neural representations. After babbling, the model can be presented with new sound samples to learn, and after a few practice attempts the model is capable of producing the sound in a feedforward manner.

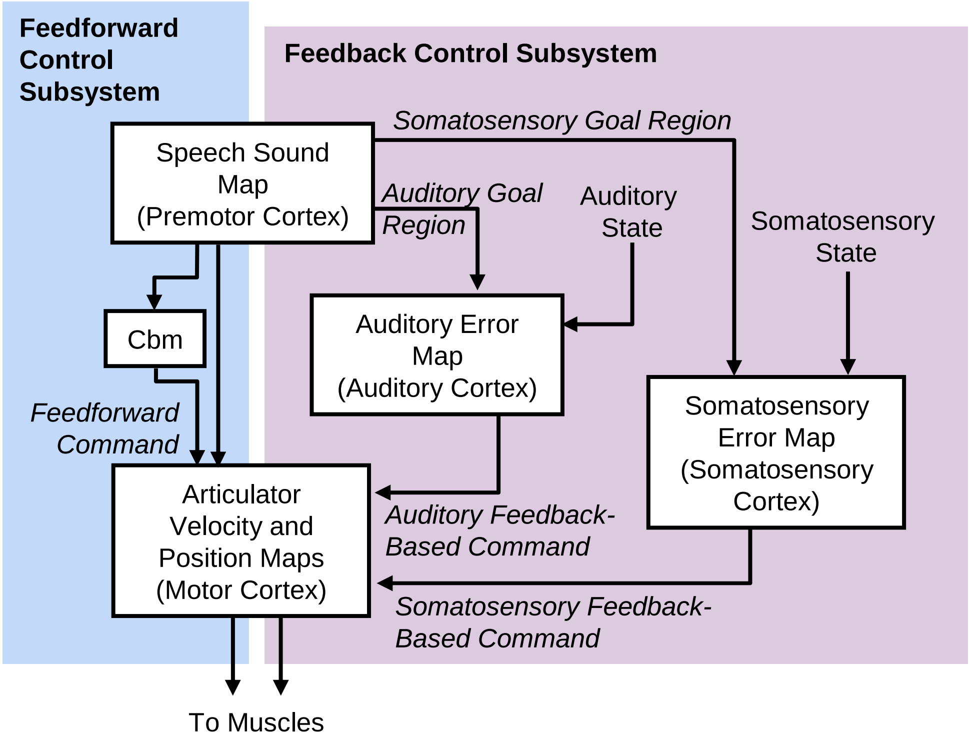
Feedforward Control Subsystem
Feedback Control Subsystem
Somatosensory Goal Region
Speech Sound Map (Premotor Cortex)
Auditory Goal Region
Auditory State
Somatosensory State
Cbm
Auditory Error Map (Auditory Cortex)
Somatosensory Error Map (Somatosensory Cortex)
Feedforward Command
Articulator Velocity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

Figure 1. Overview of the DIVA model. Boxes indicate maps of neurons; arrows indicate synaptic projections between neural maps. The model's components constitute two movement control subsystems: a feedforward control subsystem (blue) and a feedback control subsystem (violet).

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 lie in ventral lateral premotor cortex, and correspond to "mirror neurons" that have been identified in/near this area (e.g., Rizzolatti et al., 1996a,b, 1997). 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 (also referred to as "goals" or "targets") 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 transformed into 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 is learned over time by averaging the motor commands from previous attempts to produce the sound.

Before an infant 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. Furthermore, as the speech articulators grow, the feedback 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 (e.g., Callan et al., 2000).

The next section provides an example computer simulation in which the model learns to produce a new speech sound after babbling has been completed.

SIMULATION EXAMPLE: LEARNING FEEDFORWARD COMMANDS WITH PRACTICE

In this simulation, the model was presented with a new sound token, from an adult male speaker saying "good doggie", to learn. First a spatiotemporal auditory target region (specifying ranges of the first three formants for each time point) was formed based on the sample sound token, whose spectrogram is presented in the top panel of Figure 2. This target region was then used in successive attempts to produce the sound. Initially, the model's feedforward commands for the sound are inaccurate, and the model is forced to attempt to produce the sound under feedback control, leading to a poor production (Attempt 1 in Figure 2). With successive attempts, the feedforward command becomes more and more accurate, eventually leading to accurate production of the sound, as exemplified by Attempt 9 in Figure 2.

In the following sections, two theoretical characteristics of the model are discussed, and the results of associated experiments are presented. The two theoretical characteristics are the model's use of auditory target regions and its ability to achieve stable acoustic results using motor equivalent tradeoffs between articulatory gestures.

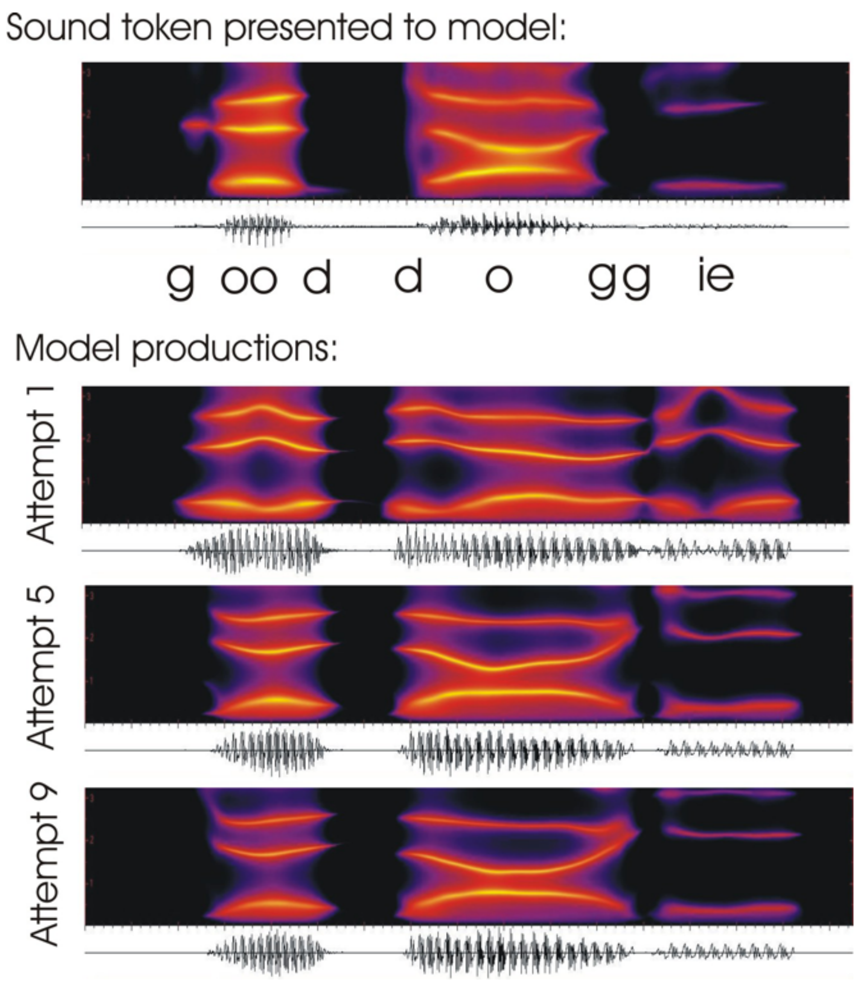

Figure 2. Spectrograms showing the first three formants of the utterance "good doggie" as produced by an adult male speaker (top panel) and by the model (bottom panels). The model first learns an auditory target for the utterance based on the sample it is presented (top panel). Then the model attempts to produce the sound, at first primarily under feedback control (Attempt 1), then with progressively improved feedforward commands supplementing the feedback control. By the 9th attempt the feedforward control signals are accurate enough for the model to closely imitate the formant trajectories from the sample utterance.

AUDITORY TARGET REGIONS – THEORY AND EXPERIMENTAL RESULTS

According to the DIVA model, the primary production target for a speech sound is a (usually time-varying) region in auditory perceptual space. Figure 3 schematizes target regions for the vowels /i/ and /e/ in a two-dimensional formant frequency space (F1,F2) for a fast speaking condition (large circles) and a clear speaking condition (small circles). (This example assumes static, rather than time-varying, targets for the vowels for simplicity.) According to the model, clear speech involves a relatively small target region for a sound as compared to fast speech. This leads to a larger contrast distance between produced sounds from neighboring categories in clear speech as compared to fast speech (see Figure 3). Shrinking of the target region can be carried out in the model through adjustment of a single parameter (see Guenther, 1995).

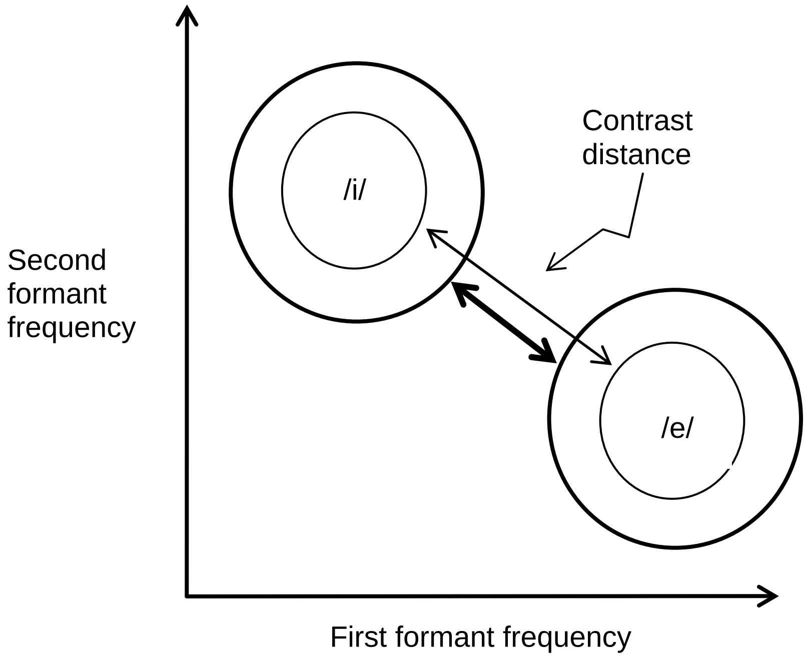

Figure 3. Schematic of auditory target regions for the vowels /i/ and /e/ in F1/F2 space. Large circles represent target regions for rapid speech, and the bold double-arrow indicates the corresponding contrast distance. Small circles indicate target regions for clear speech, and the thin double-arrow indicates the corresponding contrast distance. In going from fast to clear speech, the auditory target regions shrink and contrast distance increases accordingly.

According to the model, the auditory perceptual target for a sound is learned by listening to examples of that sound spoken by other speakers (e.g., the parents of a child), as well as monitoring good self-productions of the sound. An issue raised by this view of auditory target region learning is the following: does an individual's perceptual acuity affect the size of that individual's target regions for speech production? One might imagine that an individual with good perceptual acuity (i.e., a better than average ability to discriminate different samples of the same speech sound) might be more particular about which sounds he/she considers to be good examples of a sound category. This would be reflected by smaller target regions for the sound, which in turn would be reflected in the individual's productions by greater contrast distance between that sound and neighboring sounds.

We have tested this hypothesis in experiments involving measurement of an individual's perceptual acuity for a sound continuum between two similar sounds (e.g., “who'd” vs. “hood”), and measurement of contrast distance for the individual's productions of the sounds that form the endpoints of the continuum (Perkell et al., submitted; Perkell et al., in press). Results from these experiments are summarized in Figure 4. In general, speakers with higher perceptual acuity tend to produce larger contrasts, suggesting that they utilize smaller auditory target regions for speech production. A second general trend is for contrast distance to increase as speaking rate is decreased or clarity is increased, in keeping with the shrinking of target regions for slower/clearer speech as implemented in the DIVA model (Guenther, 1995).

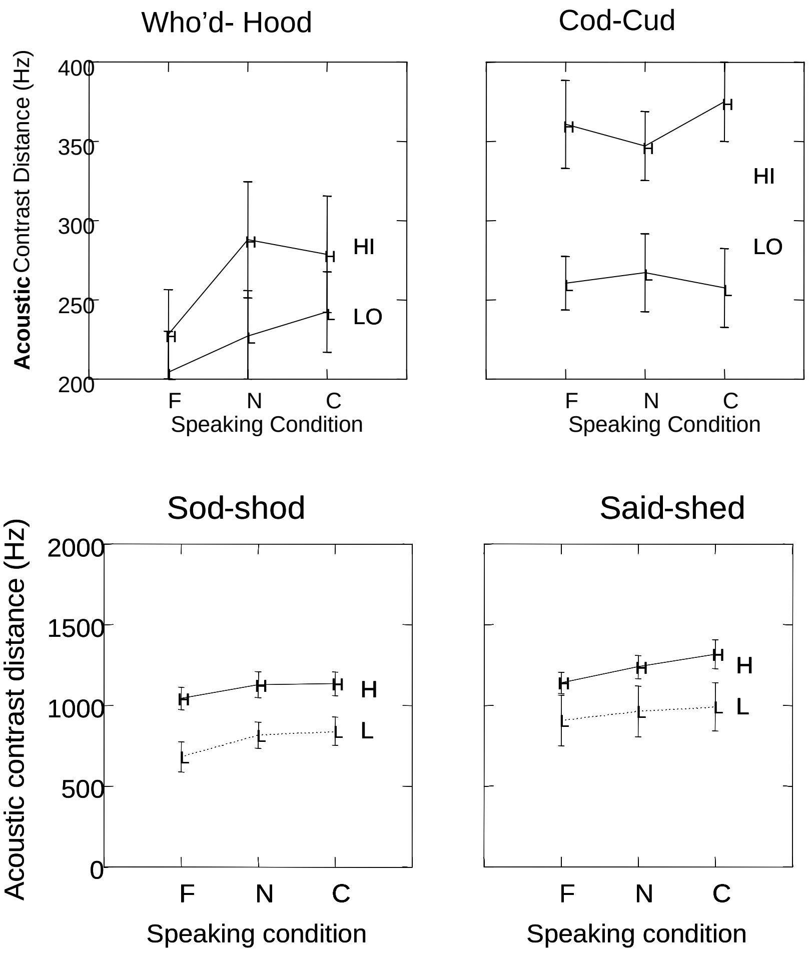

Figure 4. Summary of experimental results indicating a relationship between speaker perceptual acuity and produced contrast distance for four different contrasts: "who'd-hood" (top left), "cod-cud" (top right), "sod-shod" (bottom left), and "said-shed" (bottom right). Curves labelled "HI" (or "H") correspond to speakers with relatively high perceptual acuity for sounds on a continuum between the two contrasted sounds (e.g., a "who'd-hood" continuum for the top left plot). Curves labelled "LO" (or "L") correspond to speakers with lower acuity. "F", "N", and "C" refer to fast, normal, and clear speaking conditions.

MOTOR EQUIVALENCE – THEORY AND EXPERIMENTAL RESULTS

Because the primary target for speech production in the DIVA model is an auditory target region, and because the model uses a directional mapping between planned auditory trajectories and articulator movements, the vocal tract shapes used to produce a particular speech sound in different phonetic contexts or across repetitions can vary. Specifically, there will be articulatory trading relations in which the positions of two different articulators will covary in a systematic way that maintains stability of the acoustic signal while allowing variation of the individual articulator positions. This flexibility in choosing the articulator configuration to produce a sound allows for more efficient movements to that sound from different phonetic contexts and thus allows the speaker to produce the sound with minimal articulatory effort (Guenther, 1994, 1995; Guenther et al., 1998; see also Lindblom, 1983, 1996).

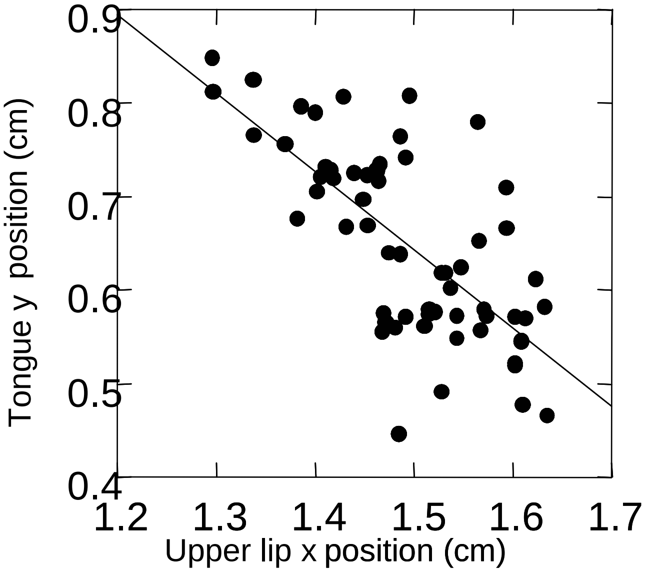

Figure 5. Articulatory trading relations between lip protrusion (x axis) and tongue body height (y axis) when a speaker produces multiple repetitions of the phoneme /u/.

Using electromagnetic midsagittal articulometry (EMMA), we have identified this type of articulatory trading relationship during speech production, including production of the American English phonemes /u/ (Perkell et al., 1993) and /r/ (Guenther et al., 1999). Figure 5 illustrates a systematic tradeoff between upper lip protrusion (x axis) and tongue body height (y axis) during multiple repetitions of /u/. Raising the tongue body and protruding the lips both have the acoustic effect of lowering F1, and individuals use the two articulations to different degrees in different repetitions while maintaining stable acoustic results. Figure 6 illustrates the use of trading relations between the length/degree of the tongue constriction and the length of the front cavity in two speakers producing /r/ (whose primary acoustic cue is a low F3; e.g., Stevens, 1998) in different phonetic contexts. When producing an /r/ that is preceded by /g/, it is easy to achieve a long and narrow tongue constriction (which has the effect of producing a low F3) by slightly lowering the tongue from its /g/ position, as seen in the red outlines in the figure. When producing the /r/ after /a/, on the other hand, it is easier to form a long front cavity by raising the tongue tip (which also has the acoustic effect of lowering F3) than it is to create a long and narrow constriction; speakers accordingly use a tongue tip raising gesture to produce /r/ after /a/, as indicated by the blue outlines in the figure. Thus, by using an auditory target rather than an articulatory target for /r/ across contexts, speakers are able to minimize the amount of movement required to produce a desired acoustic effect.

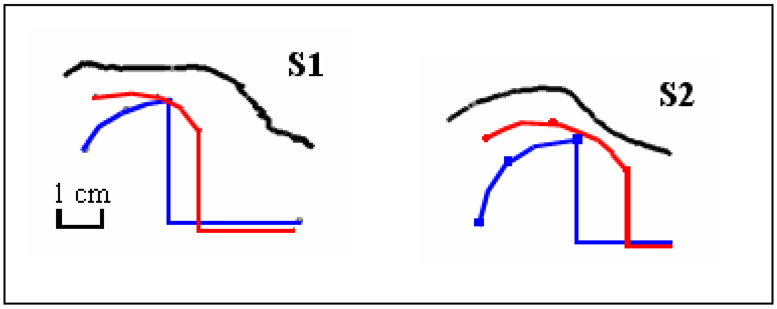

Figure 6. Trading relations between tongue body constriction length/degree and the length of the front cavity during American English /r/ production in two experimental subjects. Black lines indicate palate; red lines indicate tongue shape for /r/ after /g/; blue lines indicate tongue shape for /r/ after /a/. The lips are to the right.

CONCLUDING REMARKS

The model described herein has been used to address a wide range of EMG, kinematic, and acoustic data concerning the production of speech sounds, including observations of motor equivalence, coarticulation, contextual variability, speaking rate effects, and perception-production interactions (e.g., Guenther, 1994, 1995; Guenther et al., 1998, 1999; Perkell et al., 1993; Perkell et al., submitted; Perkell et al., in press). The model has also been used to address the development of speaking skills (e.g., Guenther, 1995; Callan et al., 2000). Because the model is formulated as a neural network whose components can be easily interpreted in terms of regions of the human brain, it is also a straightforward matter to use the model to interpreting functional neuroimaging results, and to make predictions that can be tested with neuroimaging (e.g., Guenther et al., 2003). The model thus provides a unified and detailed account of the brain functions underlying the development of speaking skills and the control of movements of the speech articulators.

ACKNOWLEDGEMENTS

This research was supported by the National Institute on Deafness and other Communication Disorders.

REFERENCES

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

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. & Ghosh, S.S. (2003). A model of cortical and cerebellar function in speech. *Proc. XVth International Congress of Phonetic Sciences*, Barcelona, 169-173.

Guenther, F.H., Ghosh, S.S., & Nieto-Castanon, A. (2003) A neural model of speech production, *Proceedings of the 6th International Seminar on Speech Production*, Sydney, 85-90.

Lindblom, B. (1983) Economy of speech gestures, In *The Production of Speech* (edited by P. MacNeilage), New York: Springer-Verlag, 217-245.

Lindblom, B. (1996). Role of articulation in speech perception: clues from production, *Journal of the Acoustical Society of America*, *99*,1683-1692.

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 (edited by W. J. Hardcastle & A. Marchal), Boston: Kluwer Academic Publishers, 131-149.

Perkell, J.S., Matthies, M.L., Svirsky, M.A. and Jordan, M.I. (1993). Trading relations between tongue-body raising and lip rounding in production of the vowel /u/: A pilot motor equivalence study, *Journal of the Acoustical Society of America, 93*, 2948-2961.

Perkell, J.S., Guenther, F.H., Lane, H., Matthies, M. L. Stockmann, E., Tiede, M. & Zandipour, M. (submitted). The distinctness of speakers' productions of vowel contrasts is related to their discrimination of the contrasts. *Journal of the Acoustical Society of America*.

Perkell, J.S., Matthies, M.L., Tiede, M., Lane, H., Zandipour, M., Marrone, N., Stockmann, E., & Guenther, F.H. (in press). 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*.

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

Rizzolatti, G., Fadiga, L., Matelli, M., Bettinardi, V., Paulesu, E., Perani, D., & Fazio, F. (1996b) Localization of grasp representations in humans by PET: 1. observation versus execution, *Experimental Brain Research*, *111*, 246-252.

Rizzolatti, G., Fogassi, L., & Gallese, V. (1997) Parietal cortex: from sight to action, *Current Opinion in Neurobiology*, *7*, 562–567.

Stevens, K. N. (1998) *Acoustic Phonetics*, Cambridge, MA: MIT Press.