Acoustic Space Movement Planning in a Neural Model of Motor Equivalent Vowel Production

Dave Johnson and Frank H. Guenther*

Department of Cognitive and Neural Systems, Boston University, Boston, MA 02215

ABSTRACT

Recent evidence suggests that speakers utilize an acoustic-like reference frame for the planning of speech movements. DIVA, a computational model of speech acquisition and motor equivalent speech production, has previously been shown to provide explanations for a wide range of speech production data using a constriction-based reference frame for movement planning. This paper extends the previous work by investigating an acoustic-like planning frame in the DIVA modeling framework. During a babbling phase, the model self-organizes targets in the planning space for each of ten vowels and learns a mapping from desired movement directions in this planning space into appropriate articulator velocities. Simulation results verify that after babbling the model is capable of producing easily recognizable vowel sounds using an acoustic planning space consisting of the formants F1 and F2. The model successfully reaches all vowel targets from any initial vocal tract configuration, even in the presence of constraints such as a blocked jaw.

1.0 Introduction

It is useful to think of speech production as the process of forming a trajectory in some planning space, or reference frame, so that the trajectory passes through a sequence of targets, each corresponding to a different phoneme in a phoneme string. There are many different forms that the planning reference frame might take. Several recent models have used reference frames that correspond to the locations and degrees of certain key constrictions in the vocal tract. The task-dynamic model (Saltzman and Munhall, 1989) and DIVA model (Guenther, 1994; 1995) use constriction-based planning spaces and are capable of motor-equivalent speech production. Recent evidence suggests, however, that humans use a planning space that is more closely related to acoustic parameters. For example, Perkell, Matthies, Svirsky, and Jordan (1993) studied production of the vowel /u/ and hypothesized that "[t]he objective of articulatory movements is an acoustic goal", rather than a goal more closely related to the articulators such as a constriction goal, based on experimental results indicating that speakers use trade-offs in constriction parameters (lip rounding and tongue-body raising) to reach an acoustic goal such as a target value of the second formant frequency (F2). Analogous results have recently been observed for consonant production (Perkell, Matthies, and Svirsky, 1994). These results suggest that speakers are not planning movements to constriction targets, but instead are planning movements toward acoustic targets. This in turn suggests that speech movements are planned in a more acoustic-like reference frame. This makes sense since the true goal of the speech production system is the creation of an acoustic signal that can be properly interpreted by listeners, not the production of specific constrictions in the vocal tract.

Guenther (1994; 1995) describes a self-organizing neural network model of speech acquisition and production called DIVA that utilizes a constriction-based reference frame for speech movement planning. Guenther (1994) demonstrated the model's ability to produce articulator movements that realize desired phoneme strings even in the presence of external perturbations or constraints applied to the articulators (e.g., complete blockage of jaw movement). The ability to use different motor means to achieve the same goal is called *motor equivalence* and is a ubiquitous characteristic of biological motor systems. As in human movements, compensation in the model is automatic; i.e., no new learning is required under the constraining conditions and compensation occurs without invoking special strategies to deal with the constraints. This work was extended in Guenther (1995), which showed how the model provides new and insightful explanations for many long-studied speech production phenomena, including contextual variability, velocity/distance relationships, speaking rate effects, carryover coarticulation, and anticipatory coarticulation.

The research described in this paper extends these prior results by investigating an acoustic-like planning space consisting of the first two formants of the speech signal in place of the constriction-based planning space used in Guenther (1994, 1995). Furthermore, the version of the model described here produces true acoustic output, which was not possible in the model of Guenther (1994; 1995) due to the simplistic articulatory structure used in those works.

2.0 Model Description

An overview of the model is shown in Figure 1. The model utilizes a babbling phase, during which synaptic weights in the adaptive neural mappings (shown as filled semicircles in Figure 1) are tuned, and a performance phase, during

* Dave Johnson supported in part by the Advanced Research Projects Agency (ONR N00014-92-J-4015) and the Office of Naval Research (ONR N00014-91-J-4100). Frank Guenther supported in part by the Air Force Office of Scientific Research (AFOSR F49620-92-J-0499).

which arbitrary phoneme strings specified by the modeler are produced as continuous movements of the speech articulators. The model represents information in three distinct reference frames: a phonetic frame, a planning frame, and an articulator frame. These frames are discussed in the following paragraphs, which describe the model components.

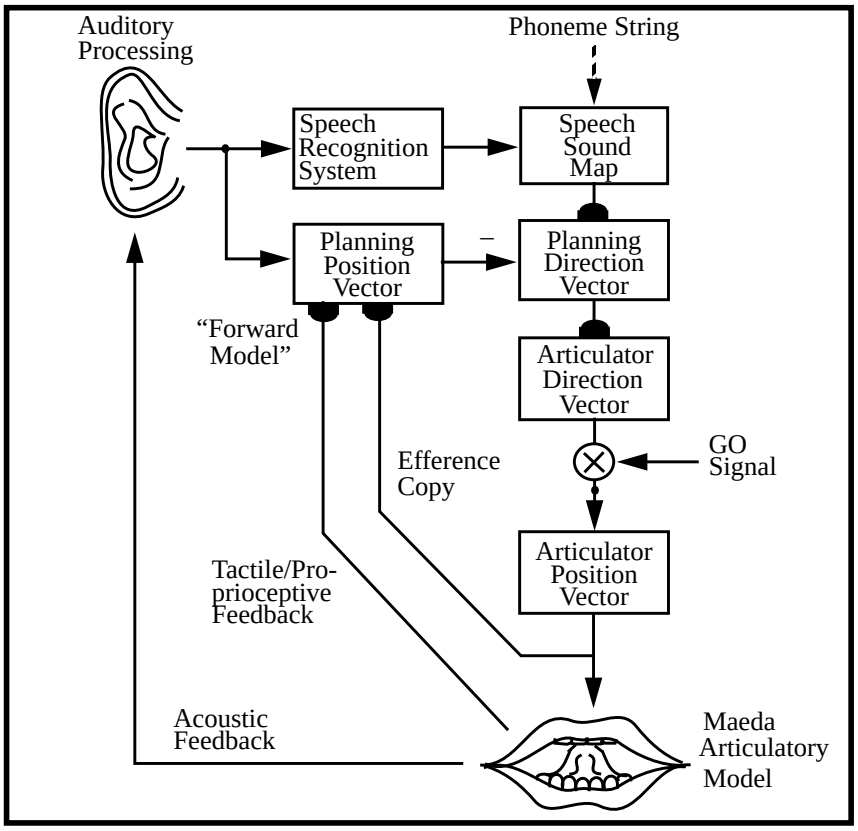

FIGURE 1. Overview of the DIVA model. Learned mappings are indicated by filled semicircles.

Speech Sound Map and Speech Recognition System. The Speech Sound Map in Figure 1 represents information in a phonetic reference frame. Each cell in this map corresponds to a different phoneme. The cell corresponding to the phoneme to be produced (or learned during the babbling phase) has an activity level of 1; all other cells in the map have zero activity. Although this paper is principally concerned with the production of vowels, consonants are also represented in the Speech Sound Map.

During babbling, the Speech Recognition System monitors the acoustic signal produced by the model (after an "auditory processing" stage that extracts formant values) and activates the appropriate cells in the Speech Sound Map when phonemes are detected. This allows learning in the weights projecting from the active Speech Sound Map cell to the cells in the Planning Direction Vector; these weights encode a target for the phoneme in planning coordinates. These targets take the form of *convex regions* in planning space. Guenther (1994; 1995) describes how considerations of speech motor development in infants suggest that sound targets take the form of convex regions, rather than points, in planning space. Guenther (1995) goes on to show how convex region targets can provide intuitive and elegant explanations for many speech production phenomena that were previously studied using point target models. The reader is referred to those works for further discussion of this topic, which is beyond the scope of the current paper.

Planning Direction Vector and Planning Position Vector. In the current simulations, the model's planning space is the set of all possible combinations of the formants F1 and F2. (Although the model computes F3 and uses it to drive the speech synthesizer, F3 is not currently used in the learning process.) As discussed in the introduction, this planning space replaces the constriction-based planning space used in Guenther (1994; 1995). The Planning Position Vector stage represents the current state of the vocal tract within the planning reference frame. This is used to calculate the desired movement direction, which is formed by subtracting the Planning Position Vector from the current sound's target at the Planning Direction Vector stage. It is hypothesized that humans have access to at least three types of information that convey the state of the vocal tract within the planning reference frame. Auditory information can provide formant values from a self-generated acoustic signal, but auditory feedback is too slow to be useful in the control of ongoing speech. More likely sources are motor command efference copy and tactile/proprioceptive feedback information. This information can be used to form a *forward model* (e.g., Jordan and Rumelhart, 1992) that maps articulator and vocal tract information into the formant values that result from the current shape of the vocal

tract. The forward model is schematized by the filled semicircles at the Planning Position Vector block in Figure 1, and is currently computed off-line. Future simulations will incorporate forward model learning into the babbling phase used to train the other learned mappings in the model.

The Planning Direction Vector is computed simply by taking the difference between the current sound's target (available through the adaptive weights projecting from the Speech Sound Map to the Planning Direction Vector) and the current configuration of the vocal tract represented in planning coordinates (available from the Planning Position Vector). For example, the Planning Direction Vector during production of a vowel might correspond to something like "lower F1 and raise F2". This vector of activities is then mapped into a set of articulator movements that carry out the desired formant changes via the adaptive weights projecting from the Planning Direction Vector to the Articulator Direction Vector.

Articulator Direction Vector and Articulator Position Vector. These neural vectors represent information within the articulator reference frame. DIVA uses an articulatory model of the vocal tract derived from the principal components analysis of cineradiographic and labiofilm data from French talkers (Maeda, 1990). The Maeda articulatory model defines seven shape parameters, or articulatory degrees of freedom (DOFs): (1) jaw height, (2) tongue-body position, (3) tongue-body shape, (4) tongue-tip position, (5) lip height (aperture), (6) lip protrusion, (7) larynx height. The seven-dimensional articulator space is the set of all possible 7-tuples of Maeda articulator values, and each vocal tract configuration corresponds to exactly one point in this articulator space. Each Maeda articulator takes on a real value in the interval [-3, 3] and may be regarded as a coefficient that weights an eigenvector. The sum of these weighted eigenvectors is a vector of points in the midsagittal plane that defines the outline of the vocal tract shape. The resulting vocal tract shape is transformed into an area function which is then processed to obtain acoustic output and spectral properties of the vocal tract during speech. Acoustic output is produced using a Klatt-based formant synthesizer.

The Articulator Direction Vector represents the desired movement direction in articulator space. Cells in the Articulator Position Vector integrate these activities (after multiplicative gating by a GO signal which controls movement speed) to produce position commands for the seven articulators.

The DIVA Babbling Phase. Babbling in the model is produced by inducing movements of the speech articulators by randomly activating the Articulator Direction Vector cells, which leads to movements of the speech articulators. Tactile and proprioceptive feedback provides information about the changing shape of the vocal tract within the planning reference frame (through the forward model), and acoustic feedback processed by the speech recognition system provides phonetic information. The combination of articulatory information (in the form of the randomly activated movement commands) and planning space information from the forward model allows tuning of the mapping between the Planning Direction Vector and the Articulator Direction Vector. The tuning process can be thought of as learning which articulator movements will move the vocal tract in a desired direction in planning space so as to allow the articulators to later carry out planned trajectories. The combination of phonetic information from the speech recognition system and planning space information from the forward model allows tuning of the mapping between the Speech Sound Map and the Planning Direction Vector. This tuning process can be thought of as learning a target in planning space for each speech sound. When a sound is babbled, the sound's target is modified based on the position in planning space that led to production of the sound.

The DIVA Performance Phase. After babbling, the model can articulate arbitrary phoneme strings using the set of learned phonemes in any combination. The version of the model that used a simplified articulatory structure (Guenther, 1994; 1995) could produce arbitrary combinations of a set of 29 phonemes, including both vowels and consonants. Because the current version of the model does not yet learn consonants, only the ten learned vowels can currently be combined to form phoneme strings.

Performance of a phoneme string can be visualized as follows. The Speech Sound Map cell corresponding to the first phoneme in the string is activated. This cell's activity propagates through the weights projecting to the Planning Direction Vector, effectively "reading out" the phoneme's learned target. The Planning Direction Vector represents the difference between this target and the current state of the vocal tract; in other words, the Planning Direction Vector codes the desired movement direction in planning space. This is then mapped into an appropriate set of articulator velocities through the learned mapping from the Planning Direction Vector to the Articulator Direction Vector. As the articulators move, the shape of the vocal tract, registered through tactile and proprioceptive feedback at the Planning Position Vector stage, gets closer and closer to the target for the speech sound. This causes the Planning Direction Vector activity to get smaller and smaller, leading to a slowing and stopping of articulator movements as the target is reached. These processes are carried out automatically by the temporal dynamics of the neural network. The time course of activity of the Planning Direction Vector cells can be thought of as the planned trajectory in acoustic coordinates. When Planning Direction Vector activity is sufficiently close to zero (i.e., when the sound has been completed), the Speech Sound Map cell corresponding to the next phoneme in the string is activated, and the process repeats. The result is a time course of articulator positions that can be viewed as a real-time animation sequence on a computer monitor.

3.0 Simulation Results

Simulations of the model were carried out on a Sparc-10 workstation. T en English vowels were learned during babbling. Synthesis of the the model's vocal tract configurations while producing each vowel in isolation resulted in easily recognized vowel sounds. Each vowel can be produced by the model from any starting configuration of the vocal tract. As illustrated in Figure 2, the resulting vocal tract shapes correspond roughly to shapes seen in humans producing the same vowels, even though no vocal tract shape information is encoded in the targets learned by the model.

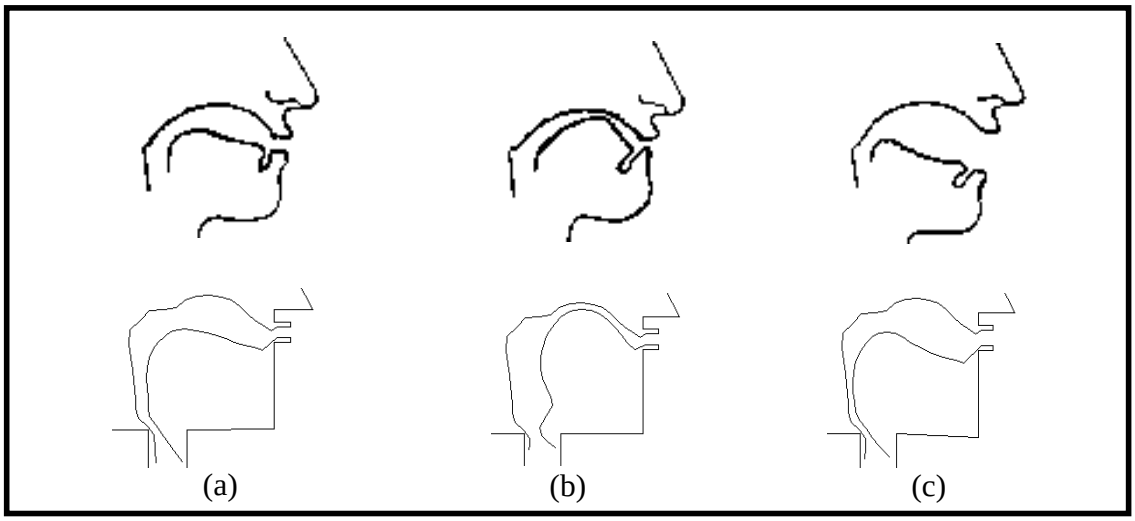

FIGURE 2. Vocal tract configurations corresponding to different vowels. The top row shows schematics of the profiles used by humans (top row; after Flanagan, 1972) and the bottom row shows the configurations produced by the model. (a) The central vowel /ʌ/ as in "up". (b) The high fr ont vowel /i/ as in "beet". (c) The low back vowel /a/ as in "father".

Each of the ten vowels were also successfully produced with the jaw blocked at various positions, demonstrating motor equivalence. With the jaw blocked, other articulators such as the tongue compensated, allowing the vocal tract to assume an overall shape that reached the acoustic tar get for the vowel. Phonemes produced with the jaw blocked were acoustically indistinguishable from phonemes that were produced with an unconstrained jaw.

4.0 Concluding Remarks

Earlier simulations of DIV A with constriction-based planning provided explanations for many speech production phenomena (Guenther, 1995). It is expected that similar results will occur with a more acoustic-like planning space such as the one described in this paper , and future simulations will investigate this issue. Future research will also address the production of stop and fricative consonants, liquids, glides, and dipthongs.

5.0 References

Flanagan, J.L. (1972). *Speech analysis, synthesis, and perception*. New York: Springer-Verlag.

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*, in press.

Jordan, M.I., and Rumelhart, D.E. (1992). Forward models: Supervised learning with a distal teacher . *Cognitive Science*, **16**, 307-354.

Maeda, S. (1990). Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model. In Hardcastle W.J. and Marchal, A. (eds). *Speech Production and Speech Modelling*, Kluwer Academic Publishers, The Netherlands. pp. 131-149.

Perkell, J.S., Matthies, M.L., and Svirsky , M.A. (1994). Articulatory evidence for acoustic goals for consonants. *Journal of the Acoustical Society of America*, **96**(5), Pt. 2, 3326.

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.

Saltzman, E. L., and Munhall, K. G. (1989). A dynamical approach to gestural patterning in speech production. *Ecological Psychology*, **1**, 333-382.