A MODELING FRAMEWORK FOR SPEECH MOTOR DEVELOPMENT AND KINEMATIC ARTICULATOR CONTROL

Frank H. Guenther
Center for Adaptive Systems and Department of Cognitive and Neural Systems
Boston University, Boston, MA 02215

ABSTRACT

This paper presents three hypotheses that are central to a computational model of speech production: (1) Sound targets take the form of regions, rather than points, in a planning reference frame. (2) The planning frame is more acoustic-like than the frames used in most recent models. (3) A direction-to-direction mapping transforms planned trajectories into articulator movements. These hypotheses are supported by experimental data and simulation results.

1. INTRODUCTION: REFERENCE FRAMES AND MAPPINGS

It is useful to think of speech production as the process of formulating a trajectory within a planning reference frame to pass through a sequence of targets, each corresponding to a different phoneme in the string being produced. This trajectory can then be mapped into a set of articulator movements that carry out the planned trajectory. The articulator movements are defined within an articulatory reference frame that relates closely to the musculature or primary movement degrees of freedom of the speech articulators. The process of mapping from the planning frame to the articulator frame need not wait until the entire trajectory has been planned, but instead may be carried out in concurrence with trajectory planning.

This paper addresses three important questions that arise within this view of the speech production process. First, what is the nature of the phonemic targets? Second, what is the nature of the planning reference frame? Finally, what is the nature of the mapping from the planning frame to the articulator frame?

The answers provided in this paper arise from a computational model of speech production called DIVA. This model is briefly introduced in Section 2. Sections 3 through 5 then address the three questions posited above. Simulation results verifying the model's ability to produce vowels are presented in Section 6. (More thorough simulations of an earlier version of the model are described elsewhere [1], [2].) Finally, Section 7 shows how the model's answers to the questions posed above lead to a simple explanation for the anomalous observation that the same speaker will often use entirely different vocal tract configurations to produce the sound /ɪ/ in different contexts.

2. MODEL DESCRIPTION

An overview of the DIVA model is shown in Figure 1. The model is formulated as a self-organizing neural network that undergoes 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 which arbitrary phoneme strings specified by the modeler are produced as continuous movements of the speech articulators. The main components of the model are briefly described in the following paragraphs; more complete descriptions are given elsewhere [1], [2].

Each cell in the Speech Sound Map in Figure 1 corresponds to a different phoneme. The cell corresponding to the phoneme to be produced has an activity level of 1; all other cells in the map have zero activity. 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; this target can later be used to produce the sound. The nature of these targets is the subject of Section 3.

The Planning Position Vector stage represents the current state of the vocal tract within the planning reference frame. This is needed to calculate the desired movement direction, which is formed at the Planning Direction Vector stage by subtracting the Planning Position Vector from the current sound's target. The time course of activity of the Planning Direction Vector constitutes the planned trajectory, specified within the planning reference frame described in Section 4.

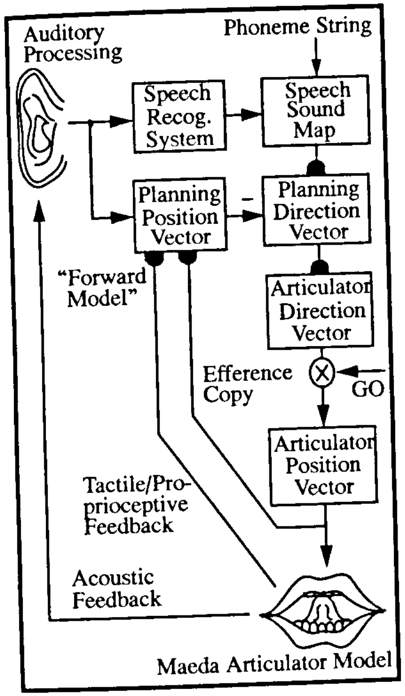

FIGURE 1. Overview of the DIVA model. Learned mappings are indicated by filled semicircles.

The desired trajectory in planning coordinates is mapped into a set of articulator movements via the adaptive weights projecting from the Planning Direction Vector to the Articulator Direction Vector. This mapping from the planning frame to the articulator frame is the subject of Section 5.

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 that controls movement speed) to produce articulator position commands. The current version of DIVA uses an articulator model created by Maeda [3]. This model consists of seven variables (jaw height, tongue body position, tongue body shape, tongue tip position, lip height, lip protrusion, larynx height) that make up the seven dimensions of the articulator reference frame.

The current version of the model includes two refinements to the version described in [1], [2]. First, the Maeda articulator set replaces the simplified articulatory structure of the earlier version, allowing synthesis of an acoustic signal based on vocal tract shape. Second, the current version uses a more acoustic-like planning reference frame. This is discussed further in Section 4.

3. PHONEMIC TARGETS

To explain how infants learn language-specific and phoneme-specific limits on variability in the production of speech sounds, Guenther [1], [2] posited a convex region theory for the targets of speech. Within this theory, the target for a speech sound consists of a range of acceptable values for each dimension of the planning frame. For example, the target for a vowel consists of a small range of acceptable values along the acoustically important dimension of tongue body constriction degree, and a much larger range along the less important dimension of velum height. This is in contrast to traditional targets, which are "all or nothing" in nature; that is, a sound's target either specifies a single target value along a dimension, or the dimension is completely unspecified.

As discussed in [2], convex region targets provide an intuitive explanation for contextual variability. The specific position within the target region that will be reached by the model during production depends on factors such as context and speaking rate; therefore, the model may end up on any point within the target region. Experimental results indicate that speakers tend to produce more variation along acoustically unimportant dimensions than they do along acoustically important dimensions [4]. This is explained by the model since the target ranges along important dimensions are much smaller than the ranges along unimportant dimensions.

Further implications of the convex region theory on several long-studied speech production phenomena were also investigated in [2]. First, it was shown that this theory provides a parsimonious explanation for a collection of speaking rate effects not previously treated by a single model. This treatment included an explanation for the experimental observation that increased speaking rate can lead to increased velocities for consonant movements but decreased velocities for vowel movements [5]. The model explains this behavior even though both sound types are produced by the same control scheme, with the effect arising from differences in the shapes of target regions for vowels and consonants.

Next, it was shown how the convex region theory provides insight into the mechanisms of carryover coarticulation. The dynamics of the neural network produce movement trajectories from the current position of the vocal tract to the closest point on the convex region target. For example, when producing the word "luke", the vocal tract configuration will move from the configuration for the back vowel /u/ to the nearest point of the target for /k/. Because the target for /k/ allows for a large amount of anterior-posterior variation in tongue body position, this will lead to a /k/ configuration that has the tongue in a relatively posterior position. For "leak", however, the tongue starts out further forward for the front vowel /i/, leading to a /k/ configuration that has a more anterior position. Thus the configuration used to produce /k/ reflects aspects of the configuration used to produce the preceding vowel, which is carryover coarticulation.

Finally, a preliminary study of anticipatory coarticulation was carried out within the framework of convex region targets. A generalization of the well-known look-ahead model of anticipatory coarticulation (e.g., [6]) was defined to allow for convex region targets and was shown to account for data not treated by the traditional look-ahead model. Because the generalized look-ahead approach posits that the amount of coarticulation is limited by the size of the convex region targets, it accounts for experimental results showing decreased coarticulation in cases where smaller targets are necessitated, including speech in languages with more crowded vowel spaces [7] and hyperarticulated speech for stress [8]. The model can also account for vowel-to-vowel anticipatory coarticulation, which is not explained by traditional look-ahead models.

4. MOVEMENT PLANNING SPACE

There are many different forms that the planning reference frame might take. For example, MacNeilage hypothesized that the target for a speech sound is a set of muscle lengths that place the speech articulators in a particular configuration that results in the target sound [9]. Producing a string of speech sounds then involves the formation of a trajectory in muscle length space that sequentially passes through the muscle length targets corresponding to the string of sounds. However, this theory runs into problems when considering speech motor equivalence; that is, speakers can use many different muscle length trajectories to produce the same phoneme. For example, speakers can immediately produce a vowel with a bite block that holds the jaw at an unnatural position, even though proper production of the vowel in this case requires a completely different set of muscle lengths than are needed under normal conditions [10]. This suggests that the target for the vowel is specified in a frame that relates more closely to the acoustic signal than to a particular configuration of the articulators.

More recently, modelers have utilized planning frames that relate more directly to the acoustic signal and thus overcome shortcomings of MacNeilage's proposal. The most common choice has been frames that describe the locations and degrees of key constrictions in the vocal tract (e.g., [1], [2], [11]). Such constriction-based frames are typically assumed to have fewer degrees of freedom than the articulator frame, so that any target in the constriction frame can be produced by one of infinitely many different configurations in articulator coordinates. In the case of a vowel, for example, the same target tongue body constriction could be reached with the jaw high and the tongue body low under normal conditions, or with the jaw lower and the tongue body higher if a bite block is present. Models that transform movement trajectories planned in a constriction frame into articulator movements in a way that automatically compensates for articulator constraints have proven capable of explaining much of the motor equivalence seen in speech [1], [11].

This discussion touches upon an important concept of biological movement control: maximally flexible performance is achieved if movements are planned in a reference frame that relates as closely as possible to the task space for the movement (e.g., acoustic space for speech), rather than a frame that relates closely to the articulators. The difficulty in explaining motor equivalence with MacNeilage's theory occurs because muscle length targets overspecify the shape of the vocal tract; targets that require all of the articulators to be in specific positions are not flexible enough to deal with things like bite blocks that prevent some of the articulators from reaching their commanded positions. A better speech production system will instead plan trajectories in an acoustic-like space, then map these trajectories into articulator movements. If the mapping process automatically compensates for externally imposed constraints on the articulators while nearly invariantly producing the planned trajectory (as the mapping described in Section 5 does), then the planning process is greatly simplified because it does not need to account for such constraints.

Although constriction frames are more directly related to the task space for speech production than articulator frames, they are still only approximations to the real task space for speech production. The true goal of the speech production mechanism is to produce an acoustic signal that conveys linguistic units to listeners, not to produce particular vocal tract constrictions. Once we consider that the end goals of speech production are acoustic, it becomes clear that constriction planning spaces can also overspecify the shape of the vocal tract. To see this, consider vowel production and note that it is possible to produce the same acoustic result (e.g., formant values) with different configurations in constriction space. For example, when producing the vowel /u/, lip rounding and tongue body raising have similar acoustic results: they both mainly act to lower F2 [12]. Thus, changes in the tongue body constriction can be compensated for acoustically by complementary changes in lip rounding. Early experimental results appear to support the idea that speakers utilize such trading relations when producing vowels [12], and analogous results have recently been observed for consonant productions [13]. If targets are specified as constriction locations and degrees, then this type of trading relation between constrictions could not be used because the target specifies a location and degree for both constrictions. This is analogous to the problem mentioned earlier for the model of MacNeilage [9]. There, a target specifying the positions of all articulators indeed leads to the correct acoustic signal, but such a target overspecifies the shape of the vocal tract and thus eliminates the possibility for automatic motor equivalent compensation.

Overspecifying the shape of the vocal tract not only reduces the ability to compensate for perturbations or constraints on the articulators, but it can also lead to inefficient movement sequences during normal speech. To see this, consider the extreme case where each phoneme's target specifies a position for every articulator. Moving from phoneme to phoneme then requires movement of many articulators that are not acoustically important.

Further evidence against constriction targets comes from studies of the American English phoneme /r/, which is a rare example of a phoneme that has at least two very different articulations that produce nearly identical acoustic patterns [14], [15]. Figure 4 shows two such configurations for /r/, known generally as "bunched" (Figure 4a) and "retroflex" (Figure 4c). The existence of two completely different configurations for producing the same phoneme is difficult for theories that hypothesize phonemic targets and movement planning in a constriction-based reference frame rather than a more acoustic-like frame. This is because the constriction locations and degrees used to produce the two /r/'s in Figure 4 are completely different; therefore the corresponding targets must also be completely different. This leads to an unparsimonious explanation in which an individual chooses one or the other target depending on context. Although not completely unreasonable, this explanation is not particularly elegant. A more parsimonious explanation utilizing a single target specified within an acoustic-like planning frame is given in Section 7.

In summary, from a modeling standpoint it makes great sense for the speech production system to utilize an acoustic-like space for target specification and movement planning rather than a constriction space or articulator space, and experimental evidence that human production systems indeed use such a frame is starting to accumulate. In keeping with this, the model of Figure 1 currently utilizes a planning frame whose dimensions correspond to formant values (see also [12], [16]). That is, the model plans formant trajectories to reach formant targets and maps these trajectories into articulator movements as described below. This formant frame replaces the constriction planning frame used in the earlier version of the model [1], [2].

5. MAPPING FROM PLANNING FRAME TO ARTICULATOR FRAME

Trajectories planned in an acoustic-like reference frame must be carried out by articulator movements. One possibility is to use a position-to-position mapping from planning space to articulator space; i.e., map each point in formant space directly into an articulator configuration that produces the desired formants. Another possibility is to use a direction-to-direction mapping from desired movement directions in planning space into movement directions of the articulators. With this kind of mapping, the configuration used to produce a desired set of formants will depend on factors such as starting configuration and externally imposed constraints on the articulators. It has been shown elsewhere [1], [16] that direction-to-direction mappings are capable of explaining motor equivalence data that position-to-position mappings cannot explain. Therefore, the DIVA model utilizes a direction-to-direction mapping to transform the desired formant changes represented at the Planning Direction Vector stage into articulator movements at the Articulator Direction Vector stage.

Earlier work demonstrated the model's ability to reach phoneme targets even in the presence of external perturbations or constraints applied to the articulators (e.g., complete blockage of jaw movement) [1]. As in humans, 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. Simulations reported in the next section verify the ability of the current version of the model to compensate for bite blocks during vowel production, and simulations reported in Section 7 show how the direction-to-direction mapping helps explain the variability in /r/ production described above.

6. VOWEL SIMULATIONS

The earlier version of the model [1], [2] produced arbitrary combinations of a set of 29 phonemes, including both vowels and consonants, using its simplified articulatory structure and the constriction-based planning frame. The current version, which utilizes an acoustic-like planning frame, does not yet produce consonants.

Simulations of the model were carried out on a Sparc-10 workstation. Ten English vowels were learned during babbling. After learning, synthesis of the the model's vocal tract configurations while producing each vowel in isolation resulted in recognizable vowel sounds. Each vowel could 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.

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

7. /r/ SIMULATIONS

Section 4 discussed how the use of two completely different articulator configurations for /r/ by the same speaker is troublesome for models using a constriction-based planning frame. This section describes how the use of an acoustic-like planning space and a direction-to-direction mapping from the planning frame to the articulator frame provides a simple explanation for this observation.

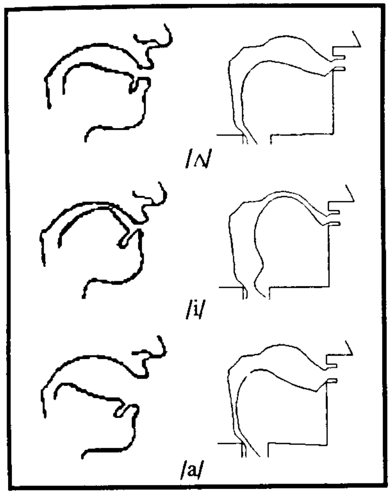

*FIGURE 2. Vocal tract configurations corresponding to different vowels. The left side shows schematics of configurations used by humans (adapted from [17]) and the right side shows the configurations produced by the model. **Top row.** The central vowel /ʌ/ as in "up". **Middle row.** The high front vowel /i/ as in "beet". **Bottom row.** The low back vowel /a/ as in "father". Model configurations approximate human configurations even though no articulatory or vocal tract shape information is used for target specification or movement planning in the model.*

It is important to note that simple target regions in acoustic space often correspond to complex regions in articulator space. The top half of Figure 3 shows a simple convex region in formant space that approximates the ranges of F1 and F2 for the phoneme /r/. The bottom half of the figure shows the corresponding region in two dimensions of articulator space. This figure was produced by fixing five of the Maeda articulators and varying the remaining two through their entire ranges to determine which configurations result in formants in the ranges specified in the top half of the figure. Note that the articulator space region is non-convex; in fact, it is broken into two distinct sub-regions. The top sub-region corresponds to a flattened tongue tip as in retroflex /r/, and the bottom sub-region corresponds to a bunched tongue configuration as in bunched /r/.

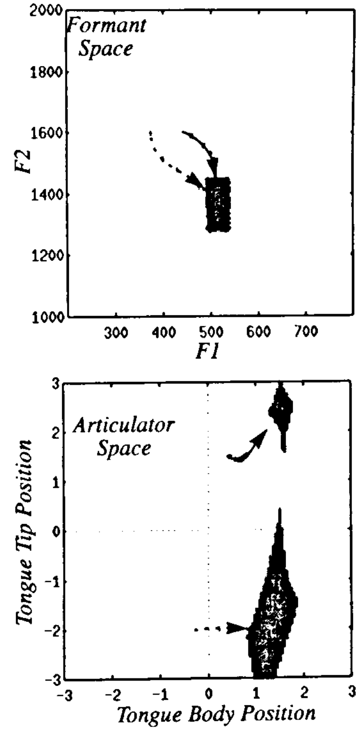

FIGURE 3. Relationship between a simple convex region corresponding to /r/ in formant space (top) and the corresponding region in articulator space (bottom). Arrows indicate model trajectories when producing /r/ starting from a /d/ configuration (solid lines) and from a /g/ configuration (dashed lines).

Assume that the same speaker is to produce the words "grab" and "drag". Often the same speaker will use a bunched /r/ following /g/ as in "grab" and a retroflex /r/ following /d/ as in "drag" [15]. The same target for /r/, consisting of the formant ranges in the top half of Figure 3, is specified to the production mechanism regardless of phonetic context. This target is manifested by the weights between the Speech Sound Map and the Planning Direction Vector stages of the model. The Planning Direction Vector activity then represents the desired movement direction in formant space. These desired formant changes are transformed into articulator movements through the direction-to-direction mapping manifested by the weights projecting from the Planning Direction Vector to the Articulator Direction Vector. These movements result in the formant trajectories shown in the top half of Figure 3. The solid arrow is the trajectory produced when /d/ precedes /r/, and the dashed arrow is the trajectory produced when /g/ precedes /r/.

The important thing to note is that the direction-to-direction mapping transforms these formant trajectories, which go to a single target region in formant space, into articulator trajectories that end up at *different* sub-regions in articulator space. The articulator space trajectories are indicated by the arrows in the bottom half of Figure 3. (Because this plot represents just two of the seven articulator dimensions, the trajectories are only approximate.) Roughly speaking, the direction-to-direction mapping causes the model to automatically move to the closest sub-region in articulator space. When /g/ precedes /r/ the bottom sub-region corresponding to bunched /r/ is closest (dashed arrow), and when /d/ precedes /r/ the upper sub-region corresponding to retroflex /r/ is closest (solid arrow). This behavior is only possible because: (i) the target is acoustic-like and does not include articulatory or constriction specifications, and (ii) formant positions in the planned trajectory are not mapped directly into articulator positions, but instead formant *changes* are mapped into articulator position *changes* by the direction-to-direction map.

The articulator configurations produced by the model in the two contexts are shown in Figure 4b,d. Although the model captures the major aspects of /r/ articulation of interest here (i.e., different sub-regions in articulator space, corresponding to a flattened tongue tip and a bunched tongue tip, are used to produce /r/ in different contexts), the model's configurations only roughly correspond to human configurations. In particular, the model's tongue tip during retroflex /r/ is not as retroflexed as a human's tongue tip. There are two reasons for this. First, the limited degrees of freedom of the Maeda articulators do not allow for much retroflexion of the tongue. Second, an important acoustic cue for /r/ is a very low F3. Because the model does not currently include F3 in the planning space, this aspect is not captured here. The sublingual cavity formed with retroflex tongue shapes is partly responsible for lowering F3. It is therefore anticipated that incorporating F3 in the planning space and using a better model of the tongue and sublingual cavity will result in /r/ productions that are more retroflexed than in Figure 4d.

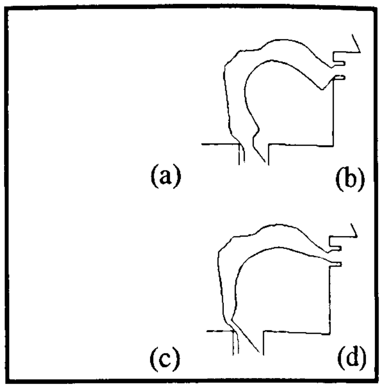

FIGURE 4. (a) Typical bunched /r/ configuration (adapted from [14]). (b) Configuration reached by the model when producing /r/ after /g/. (c) Typical retroflex /r/ configuration (adapted from [14]). (d) Configuration reached by the model when producing /r/ after /d/.

It should also be noted that the preceding phoneme is only one of the factors that affect the choice of /r/ configuration, so the model does not yet account for all aspects of /r/ variability.

8. CONCLUDING REMARKS

This paper has presented three simple hypotheses that explain a wide range of experimental data on speech articulation. These hypotheses are implemented in a computational model of speech acquisition and production. Simulation results verified the model's ability to produce vowels with or without a bite block and to explain the anomolous observation that the same speaker will often use widely different articulator configurations to produce /r/ in different contexts.

Ongoing research is addressing the addition of F3 to the planning space and the addition of consonants to the model's phonemic repertoire. It is believed that consonants, unlike vowels and /r/, may require the incorporation of constriction information in the target specification. Therefore, a hybrid planning space including both formants and constrictions will be investigated.

9. REFERENCES

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

[2] Guenther, F.H. (in press), "Speech sound acquisition, coarticulation, and rate effects in a neural network model of speech production." *Psychological Review.*

[3] 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*, Boston: Kluwer Academic Publishers, pp. 131-149.

[4] Perkell, J.S., & Nelson, W.L. (1985), "Variability in production of the vowels /i/ and /a/." *Journal of the Acoustical Society of America*, vol. 77, pp. 1889-1895.

[5] Gay, T., Ushijima, T., Hirose, H., & Cooper, F.S. (1974), "Effects of speaking rate on labial consonant-vowel articulation." *Journal of Phonetics*, vol. 2, pp. 47-63.

[6] Kozhevnikov, V.A., & Chistovich, L.A. (1965), *Speech: Articulation and perception.* Translation by Joint Publications Research Service, Washington DC, JPRS 30543.

[7] Manuel, S.Y. (1990), "The role of contrast in limiting vowel-to-vowel coarticulation in different languages." *Journal of the Acoustical Society of America*, vol. 88, pp. 1286-1298.

[8] De Jong, K., Beckman, M.E., & Edwards, J. (1993), "The interplay between prosodic structure and coarticulation." *Language and Speech*, vol. 36, pp. 197-212.

[9] MacNeilage, P.F. (1970), "Motor control of serial ordering in speech." *Psychological Review*, vol. 77, pp. 182-196.

[10] Lindblom, B., Lubker, J., & Gay, T. (1979), "Formant frequencies of some fixed-mandible vowels and a model of speech motor programming by predictive simulation." *Journal of Phonetics*, vol. 7, pp. 147-161.

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

[12] 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*, vol. 93, pp. 2948-2961.

[13] 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*, vol. 96(5), Pt. 2, p. 3326.

[14] Delattre, P., and Freeman, D.C. (1968), "A dialect study of American r's by x-ray motion picture." *Linguistics*, vol. 44, pp. 29-68.

[15] Espy-Wilson, C., & Boyce, S. (1994), "Acoustic differences between 'bunched' and 'retroflex' variants of American English /r/." Journal of the Acoustical Society of America, vol. 95(5), Pt. 2, p. 2823.

[16] Bailly, G., Laoissière, R., & Schwartz, J.L. (1991), "Formant trajectories as audible gestures: an alternative for speech synthesis." *Journal of Phonetics*, vol. 19, pp. 9-23.

[17] Bullock, D., Grossberg, S., & Guenther, F.H. (1993), "A self-organizing neural model of motor equivalent reaching and tool use by a multijoint arm." *Journal of Cognitive Neuroscience*, vol. 5, pp. 408-435.

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

10. ACKNOWLEDGEMENTS

This research was supported in part by the Air Force Office of Scientific Research (AFOSR F49620-92-J-0499). The author would like to thank Dave Johnson for his contribution to this work.