A Neural Model of Speech Production and Its Application to Studies of the Role of Auditory Feedback in Speech

Frank H. GUENTHER[1,2] and Joseph S. PERKELL[2,1,3]
[1]*Department of Cognitive and Neural Systems, Boston University, 677 Beacon Street, Boston, MA, USA*
[2]*Speech Communication Group, Research Laboratory of Electronics, MIT, Cambridge, MA, USA*
[3]*Department of Brain and Cognitive Sciences, MIT,Cambridge, MA, USA*
guenther@cns.bu.edu; perkell@speech.mit.edu

Abstract. This paper describes a neural model of speech production and perception-production interactions. This model has been developed to account for a wide variety of experimental data, ranging from kinematic analyses of articulator movements to functional imaging studies of the human brain. Hypothesized neural correlates of the model's components have been identified to facilitate testing of model predictions with techniques such as fMRI. The model also serves as a framework for interpreting and organizing the accumulating mass of data from functional imaging studies of the human brain. According to the model, the goals of speech movements are in auditory-temporal space and the movements are planned with the use of mappings between articulations and their acoustic and auditory consequences. It is hypothesized that the mappings are acquired and maintained with the use of auditory feedback. Data are presented from studies of changes in speech that occur in response to a change in hearing status. These data provide information about the nature of the mappings and how they are used in planning speech movements.

1. The DIVA Model of Speech Production

The overall objective of our research is to model the brain activity and the motor, biomechanical and sensory processes involved in speech production. Our approach is to use a combination of computational models and to develop and test them with brain imaging, psychophysical, physiological, anatomical and acoustic data. In particular, we have developed a neural network model of speech motor skill acquisition and speech production, called the DIVA model, that explains a wide range of data on contextual variability, motor equivalence, coarticulation, and speaking rate effects (Guenther, 1994, 1995a,b; Guenther, Hampson, and Johnson, 1998; Guenther and Micci Barreca, 1997; Perkell et al., 2000). This model is schematized in Figure 1. In this chapter we provide a description of the model and its neural correlates, and we present results of studies on an important aspect of the model – the role of auditory feedback in the planning of speech movements.

Each block in Figure 1 corresponds to a set of neurons that constitute a neural representation. 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 three neural mappings, indicated by filled semicircles in the figure. These mappings are later used for phoneme production.

The synaptic weights of the first mapping, labeled "convex region targets" in the figure, encode auditory and orosensory targets for each phoneme the model learned during babbling. To explain how infants learn phoneme-specific and language-specific limits on acceptable articulatory and acoustic variability, the learned speech sound targets take the form of multidimensional regions, rather than points, in auditory and orosensory spaces. This "convex region theory" of phonemic targets as multidimensional regions provides a simple and unified explanation for many long-studied speech phenomena, including aspects of anticipatory and carryover coarticulation, contextual variability, motor equivalence, velocity/distance relationships, and speaking rate effects (Guenther, 1995a).

The second neural mapping, labeled "directional mapping" in the figure, transforms desired movement directions in auditory and orosensory spaces into movement directions in an articulator space closely related to the vocal tract musculature. This mapping is related to the Moore-Penrose (MP) pseudoinverse of the Jacobian matrix relating the auditory and articulatory spaces (in effect, the model learns an approximation of the MP pseudoinverse during babbling); the model is thus closely related to pseudoinverse-style control techniques described in the robotics literature (e.g., Liégeois, 1977). The Jacobian matrix defines the transform between changes in articulator positions and the corresponding changes in auditory parameters (e.g., changes in formant frequencies). This relationship is many-to-one; that is, many different articulator movements will lead to the same changes in the auditory parameters. To control speech movements, one needs to invert this relationship; i.e., one needs to compute articulatory changes to carry out desired changes in the auditory parameters. Since this inversion process is one-to-many, a unique inverse to the Jacobian transformation cannot be calculated. Instead, a pseudoinverse must be calculated. The Moore-Penrose pseudoinverse is the pseudoinverse that commands the smallest amount of articulatory movement that can be used to achieve the desired changes in the auditory signal. The model learns an approximation to the MP pseudoinverse and uses this to map desired auditory changes into articulatory movements. The resulting controller is capable of automatically compensating for constraints and/or perturbations applied to the articulators (Guenther, 1994, 1995a; Guenther and Micci Barreca, 1997), thus accounting for the motor equivalent capabilities observed in humans when speaking with a bite block or lip perturbation.

The third mapping, labeled "forward model" in the figure, transforms orosensory feedback from the vocal tract and an efference copy of the motor outflow commands into a neural representation of the auditory signal that corresponds to the current vocal tract shape. This forward model allows the system to control speech movements without relying on auditory feedback, which may be absent or too slow for use in controlling ongoing articulator movements.

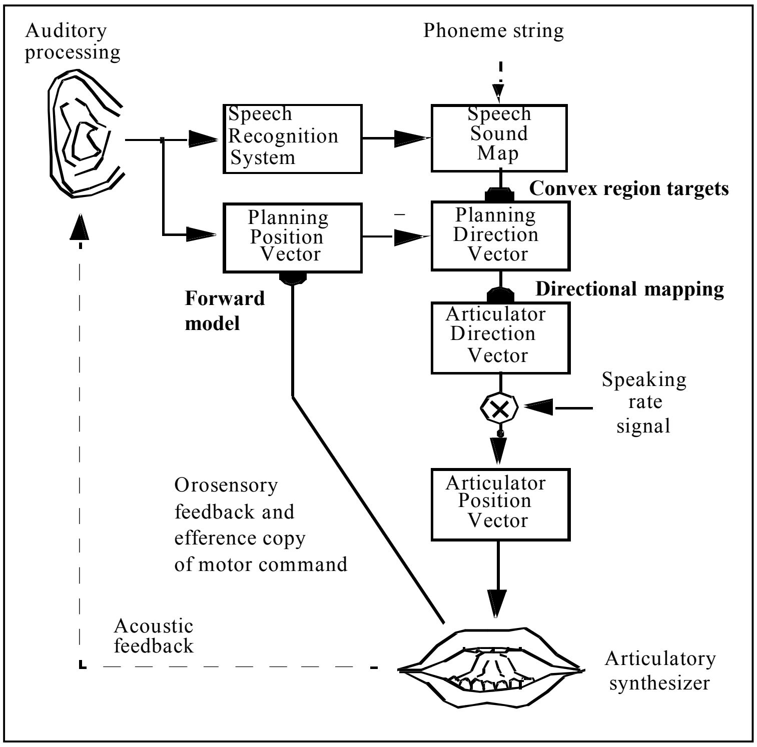

Figure 1. Overview of the DIVA model. Filled semicircles represent learned neural mappings.

Computer simulations have been used to verify that the model provides a unified explanation for a wide range of data on articulator kinematics and motor skill development (Guenther, 1994, 1995a,b; Guenther, Hampson, and Johnson, 1998; Callan et al., 2000) that were previously addressed individually rather than in a single model. The model's explanations for several speech production phenomena are discussed below, with reference to an important issue addressed by the model: the nature of the brain's "targets" for speech motor control.

The nature of speech sound targets. Most accounts of speech production involve some sort of "target" that the motor system hopes to achieve in order to produce a particular speech sound. For example, phoneme targets in the task-dynamic model (Saltzman and Munhall, 1989) take the form of locations and degrees of key constrictions of the vocal tract. Targets in the DIVA model take the form of regions in a planning space consisting of auditory and orosensory dimensions (e.g. formant ratios and vocal tract constrictions). Each cell in the model's speech sound map (see Figure 1) represents a different sound (phoneme or syllable). The synaptic weights on the pathways projecting from a speech sound map cell to cells in the planning direction vector represent a target for the corresponding speech sound in planning space. When the changing vocal tract configuration is identified by the speech recognition system as producing a speech sound during babbling, the appropriate speech sound map cell's activity is set to 1. This in turn causes learning to occur in the synaptic weights of the pathways projecting from that cell, thereby allowing the model to modify the target for the speech sound based on the current configuration of the vocal tract.

To explain how infants learn phoneme-specific and language-specific limits on acceptable articulatory variability, the targets take the form of convex regions in planning space. This "convex region theory" is a generalization of Keating's (1990) "window model" of coarticulation to a multi-dimensional movement planning space consisting of auditory and constriction dimensions in addition to articulatory dimensions (see Guenther, 1995a for further discussion of this topic). Figure 2 schematizes the learning sequence for the vowel /i/ along two dimensions of planning space, corresponding to lip aperture and tongue body height. The first time the phoneme is produced during babbling, synaptic weights that project from the speech sound map cell for /i/ are adjusted to encode the position in planning space that led to proper production of the phoneme on this trial. In other words, the model has learned a target for /i/ that consists of a single point in the planning space, as schematized in Figure 2a. The next time the phoneme is babbled, the speech sound map cell expands its learned target to be a convex region that encompasses the previous point and the new point in planning space, as shown in Figure 2b; this can occur via a simple and biologically plausible learning law (Guenther, 1995a). In this way, the model is constantly expanding its convex region target for /i/ to encompass all of the various vocal tract configurations that can be used to produce /i/.

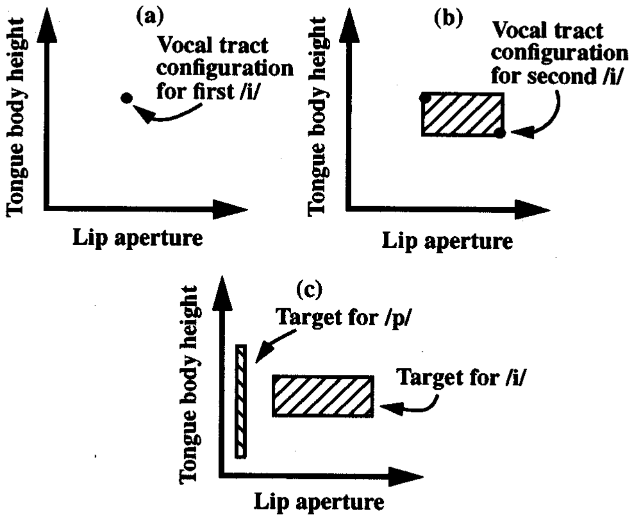

Figure 2. Learning of the convex region target for the vowel /i/ along planning dimensions corresponding to lip aperture and tongue body height. (a) The first time /i/ is produced during babbling, the learned target is simply the configuration of the vocal tract when the sound was produced. (b) The second time /i/ is babbled, the convex region target is expanded to encompass both vocal tract configurations used to produce the sound. (c) Schematized convex regions for /i/ and /p/ after many productions of each sound during babbling. Whereas the target for /i/ allows large variation along the dimension of lip aperture, the target for the bilabial stop /p/ requires strict control of this dimension, indicating that the model has learned that lip aperture is an important aspect of /p/ but not /i/.

An important aspect of this work concerns how the nervous system extracts the appropriate forms of auditory and orosensory information that define the different speech sounds. For example, how is it that the nervous system ''knows'' that it is lip aperture, and not lower lip height or upper lip height, that is the important articulatory variable for stop consonant production? How does the nervous system know that whereas lip aperture must be strictly controlled for bilabial stops, it can be allowed to vary over a large range for many other speech sounds, including not only vowels but also velar, alveolar, and dental stops? How does the nervous system of a Japanese speaker know that tongue tip location during production of Japanese /r/ can often vary widely, while the nervous system of an English speaker knows to control tongue tip location more strictly when producing /r/ so that /l/ is not produced instead?

The manner in which targets are learned in the DIVA model provides a unified answer to these questions. Consider the convex regions that result after many instances of producing the vowel /i/ and the bilabial stop /p/ (Figure 2c). The convex region for /p/ does not vary over the dimension of lip aperture but varies largely over the dimension of tongue body height; this is because all bilabial stops that the model has produced have the same lip aperture (corresponding to full closure of the lips), but tongue body height has varied. In other words, the model has learned that lip aperture is the important dimension for producing the bilabial stop /p/. Furthermore, whereas lip aperture is the important dimension for /p/, the model has learned that this dimension is not very important for /i/, as indicated by the wide range of lip aperture in the target for /i/ in Figure 2c. Finally, since convex region learning relies on language-specific recognition of phonemes by the infant, the shapes of the resulting convex regions will vary from language to language.

The convex region theory of the targets of speech provides a unified explanation for a number of long-studied speech production phenomena. A brief summary of some of these data explanations is provided below; see Guenther (1995a) for further detail.

Articulatory variability. Convex region targets provide a natural framework for interpreting data on motor variability in speech: the motor system is careful to control movements along dimensions that are important for a sound (i.e., dimensions with small target ranges), but not movements along dimensions that are not important (those with large target ranges). The model accordingly shows more variability for acoustically unimportant dimensions as compared to acoustically important dimensions. Experimental support for this comes from the study of Perkell and Nelson (1985), who found more articulatory variability along acoustically less important dimensions for the vowels /i/ and /a/. Specifically, this study showed more variability in tongue position along a direction parallel to the vocal tract midline than for the acoustically more important tongue position along a direction perpendicular to the vocal tract midline when subjects produced /i/ and /a/ sounds in different phonetic contexts and at different speaking rates.

Carryover coarticulation. The model's explanation for carryover coarticulation is simple and straightforward: when producing a phoneme from different initial configurations of the vocal tract, different positions on the convex region target will be reached, since the model moves to the closest point on the target region. The end effect of this is that the configuration used to produce a sound will depend on which sound precedes it, with the model choosing a configuration that is as close as possible to the preceding configuration. Figure 3 schematizes the situation for the target /k/ in the words "luke" and "leak". The initial front-back position of the tongue body for the preceding vowel determines the configuration of the vocal tract reached for the consonant /k/. When the back vowel /u/ precedes /k/ as in "luke", the tongue body is further back during /k/ than when the front vowel /i/ precedes /k/ as in "leak", as seen when English-speaking subjects speak these words (e.g., Daniloff et al., 1980; Kent and Minifie, 1977).

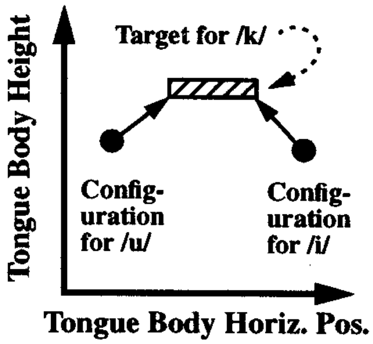

Figure 3. Convex region theory account of carryover coarticulation in /k/ production. Approaching the target for /k/ from the configuration corresponding to the back vowel /u/ in "luke" leads to a final tongue body configuration that is further back than when approaching from the configuration corresponding to the front vowel /i/ in "leak".

Anticipatory coarticulation. The model's explanation of anticipatory coarticulation posits that the target region for a speech sound is reduced in size based on context in order to provide a more efficient sequence of articulator movements. Because the amount of anticipatory coarticulation is limited by the size of the convex region targets in the model, it accounts for experimental results showing decreased coarticulation in cases where smaller targets are necessitated, including speech in languages with more crowded vowel spaces (Manuel, 1990), speech hyperarticulated for clarity (Picheney, Durlach, and Braida, 1985, 1986; Lindblom and MacNeilage, 1986), and speech hyperarticulated for stress (De Jong, Beckman, and Edwards, 1993).

2. Hypothesized Neural Correlates of the DIVA Model

One advantage of the neural network approach is that it allows one to analyze the brain regions involved in speech in terms of a well-defined theoretical framework, thus allowing a deeper understanding of the brain mechanisms underlying speech. Figure 4 illustrates hypothesized neural correlates for several central components of the DIVA model. These hypotheses are based on a number of neuroanatomical and neurophysiological studies, including lesion/aphasia studies, MEG, PET, and fMRI imaging studies, and single-cell recordings from cortical and subcortical areas in animals.

The pathway labeled ‘a’ in the figure corresponds to projections from premotor cortex to primary motor cortex, hypothesized to underlie feedforward control of the speech articulators. Pathway b represents hypothesized projections from premotor cortex (lateral BA 6) to higher-order auditory cortical areas in the superior temporal gyrus (BA 22) and orosensory areas in the somatosensory cortex (BA 1,2,3; pathway not shown in figure for clarity) and supramarginal gyrus (BA 40). These “efference copy” projections are hypothesized to carry target sensations associated with motor plans in premotor cortex. For example, premotor cortex cells representing the syllable /bi/ are hypothesized to project to higher-order auditory cortex cells; these projections represent an expected sound pattern (i.e., the auditory representation of the speaker’s own voice while producing /bi/). Similarly, projections from premotor cortex to orosensory areas in the somatosensory cortex and supramarginal gyrus represent the expected pattern of somatosensory stimulation during /bi/ production. Pathway b is hypothesized to encode the convex region targets for speech sounds in the DIVA model, corresponding to the pathway between the Speech Sound Map and Planning Direction Vector in Figure 1.

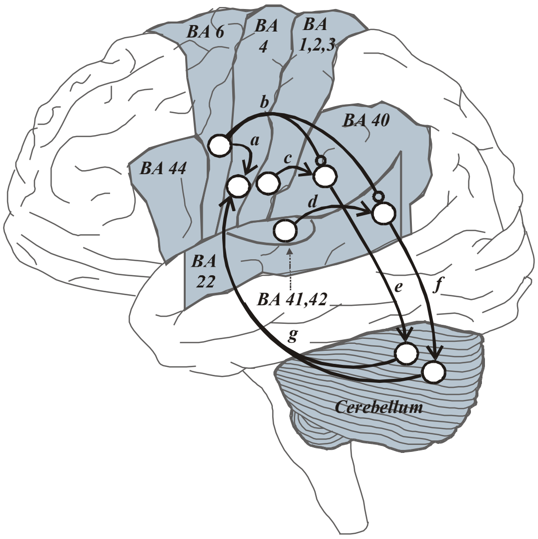

Figure 4. Hypothesized neural correlates of several central components of the DIVA model. BA = Brodmann’s Area. See text for details.

One interesting aspect of the model in Figure 4 is the role of auditory cortical areas in speech production as well as speech perception. According to the model, auditory “targets” project from premotor cortical areas to the posterior superior temporal gyrus (pathway b), where they are compared to incoming auditory information from primary auditory cortex (pathway d, corresponding to the pathway between the Planning Position Vector and Planning Direction Vector in Figure 1). The difference between the target and the actual auditory signal represents an “error” signal that is mapped through the cerebellum (pathway f), which transforms the auditory error into a motor velocity signal that can act to zero this error (pathway g; pathways f and g correspond to the Directional Mapping in Figure 1). This projection through the cerebellum to motor cortex forms a component of the Directions Into Velocities of Articulators mapping that gives the DIVA model its name. Evidence that auditory cortical areas in the superior temporal gyrus and temporal plane are involved in speech production comes from a number of neuroimaging studies. For example, Hickok et al. (2000) report activation in left posterior STG areas (planum temporale, superior temporal sulcus) during a PET visual object naming task in which the subject’s auditory feedback of his/her own productions was masked with noise. Bookheimer et al. (1995) report activations near primary auditory cortex in a similar task. Paus et al. (1996) also reported activation in the area of the left planum temporale during a PET object naming task. These authors attributed this activation to “motor-to-sensory discharges”, compatible with pathway b in Figure 4. This interpretation also receives support from a magnetoencephalography (MEG) study by Levelt et al. (1998), who showed that the auditory cortical activations during speech production slightly preceded the initiation of articulatory processes. All of these results provide support for the notion of auditory perceptual targets for speech production, in keeping with a central aspect of the DIVA model (e.g., Guenther, 1995b; Guenther et al., 1998; see also Perkell et al., 1995b; Bailly, Laboissière, and Schwartz, 1991).

The model also proposes a role for the supramarginal gyrus (BA 40) in speech production. This brain region has been implicated in phonological processing for speech perception (e.g., Caplan, Gow, and Makris, 1995; Celsis et al., 1999), as well as speech production (Geschwind, 1965; Damasio and Damasio, 1980). The current model proposes that, among other things, the supramarginal gyrus represents the difference between target oral sensations (projecting from premotor cortex via pathway b in Figure 4) and the current state of the vocal tract (projecting from somatosensory cortex via pathway c). This difference represents the desired movement direction in orosensory coordinates and is hypothesized to map through the cerebellum to motor cortex, thus constituting a second component of the Direction Into Velocities of Articulators mapping.

Not shown in Figure 4, for the sake of clarity, is the insular cortex, buried within the sylvian fissure. The anterior insula has been shown to play an important role in speech articulation (e.g., Dronkers, 1996). This region is contiguous with the frontal and central opercula, which include portions of the premotor and motor cortices related to oral movements. We adopt the view that the anterior insula has similar functional properties to the premotor and motor cortices. This view receives support from fMRI studies showing activation of anterior insula during non-speech tongue movements (Corfield et al., 1999), PET results showing concurrent primary motor cortex and anterior insula activations during articulation (Fox et al., 2001), and PET results showing concurrent lateral premotor cortex and anterior insula activations during articulation (Wise et al., 1999).

Also not shown in Figure 4 are the Forward Model pathways (see Figure 1), which are hypothesized to project from the primary motor cortex (BA 4) to the somatosensory and auditory cortical areas. These pathways are proposed to be used in place of incoming sensory information during rapid speech, when sensory feedback is too slow for the control of ongoing movements.

An important purpose of the model outlined in Figure 4 is to generate predictions that serve as the basis for focused functional imaging studies of brain function during speech. For example, the model of Figure 4 predicts that perturbation of a speech articulator such as the lip during speech should cause an increase in activation in the somatosensory cortex and supramarginal gyrus, since the perturbation will cause a mismatch between orosensory expectations and the actual orosensory feedback signal. The model further predicts that extra activation will be seen in the cerebellum and motor cortex under the perturbed condition, since pathway e in Figure 4 would transmit the extra supramarginal gyrus activation to the cerebellum and on to motor cortex (pathways e, g). We are currently testing these and other predictions of the model using fMRI and MEG.

3. Auditory feedback in adult speech production

The model in Figure 1 includes an auditory feedback pathway (left side of figure) that is responsible for the learning and maintenance of three mappings between information in different reference frames. First, articulatory commands are mapped into their expected auditory consequences (the mapping labeled Forward Model in Figure 1). Second, desired movement directions in auditory space are mapped into articulator movements (the "directional mapping" of Figure 1). These two mappings in the model are "systemic" mappings, in that they are used for the production of all speech sounds. Auditory feedback is also used to learn a third, "phoneme-specific" mapping between cells representing speech sounds and corresponding regions in auditory perceptual space (the "convex region targets" in Figure 1). The components of this mapping are specific to a particular phoneme or syllable. For example, a cell representing the phoneme /i/ in the model's speech sound map will be mapped into a target region of auditory space that corresponds to the sound /i/.

It is well known that people born deaf usually have a very difficult time learning how to speak intelligibly. On the other hand, if someone is born with hearing, learns how to speak, and then becomes deaf, that person is often able to continue speaking intelligibly for decades without being able to hear. These basic observations support the idea that learning how to speak involves establishing neural mappings such as those shown in Figure 1 under conditions of auditory feedback, and that if deafness occurs after establishment of these mappings, they can remain fairly accurate for years as long as the relationship between articulatory and auditory parameters remains constant. Thus, a full-grown adult who becomes deaf would be expected to maintain fairly accurate mappings, while a child that becomes deaf would be expected to show degradation of the mappings relating articulatory and auditory parameters due to growth-induced changes in the geometry of the vocal tract (which change the auditory-articulatory relationship).

In order to better characterize these mappings, we have been investigating the role of auditory feedback in adult speech production by observing changes in speech that occur in response to changes in hearing status, such as the loss of hearing due to disease or the acquisition of some hearing from a cochlear implant. Six of these studies are described briefly below, along with observations about how their results are related to the model.

3.1 Learned neural mappings are generally stable after onset of deafness.

The long-term stability of learned neural mappings is exemplified by the predominantly normal vowel formant patterns seen for two female CI users, FA and FB, in Figure 5. The figure shows sets of average F_1 and F_2 values (upper and lower panels) arranged by vowel for the two speakers (left and right panels). The small squares connected by dotted lines show normative values from Peterson and Barney (1952). The values indicated by unfilled circles are pre-implant, and those indicated by filled circles are (1-2 years) post-implant. The error bars indicate one standard error about the mean.

For the most part, overall vowel formant patterns (relations of formant values to one another among the vowels) appear to be relatively congruent with the normative patterns, even years after the onset of profound hearing loss (Perkell et al., 1992). The most prominent exception to this observation is for FA. Eighteen years after the onset of her profound hearing loss, pre-implant F_2 values among her front vowels /i/, /ɪ/, /ɛ/ and /æ/ were somewhat disordered with respect to the Peterson and Barney data, primarily due to relatively high values for /ɛ/ and especially /æ/. As indicated by the filled circles, after about a year with prosthetic hearing, these F_2 values are more in line with the Peterson and Barney pattern. Thus, FA's abnormal pre-implant F_2 pattern was "corrected" toward the normative pattern after some months of implant use. We hypothesize that this correction was due to a retuning of neural mappings using auditory feedback available from the cochlear implant.

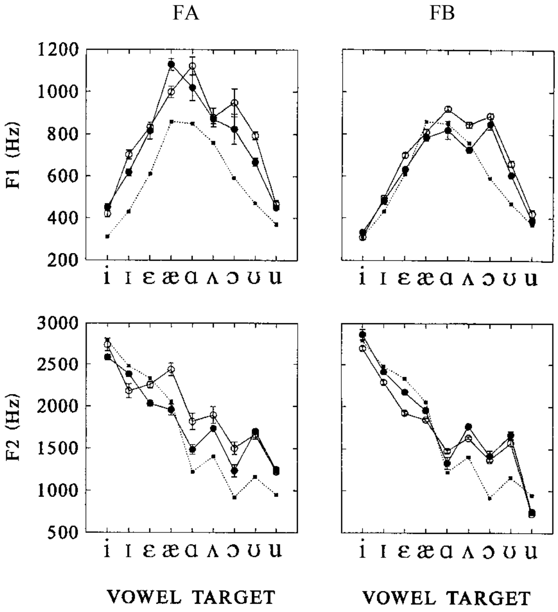

Figure 5. Average F_1 and F_2 values (upper and lower panels) arranged by vowel for two female CI users (left and right panels).

3.2 Goals for the fricative consonants /s/ and /ʃ/ are also generally stable.

Figure 6 shows values of spectral median and symmetry for /s/ and /ʃ/ produced in carrier phrases by three of five cochlear implant users studied by Matthies et al. (1994). The measurements, which reflect acoustically and perceptually salient differences between /s/ and /ʃ/, were made pre-implant, within a few months after implant and six months post-implant. Pre-implant, as exemplified by FA and MB, four of the five subjects had higher values of spectral median for /s/ than for /ʃ/, higher values of symmetry for /ʃ/ than /s/ and clear separation between the /s/ and /ʃ/ values. These results indicate a good distinction between the two consonants pre-implant – even decades following the onset of profound deafness. The good pre-implant distinctions between the sibilants in four of the subjects indicate that their systemic and phoneme-specific mappings for the production of

/s/ and /ʃ/ were generally quite stable, even in the prolonged absence of auditory feedback. On the other hand, the fifth subject (FC) had reversed values of the two measures pre-implant (consistent with the experimenters' impression that her sibilants were quite distorted), indicating an unusually extreme distortion of her neural mappings. After months of implant use, FC's spectral median and symmetry values were greatly improved. We believe this improvement resulted from a corrective retuning of the neural mappings after she received the cochlear implant.

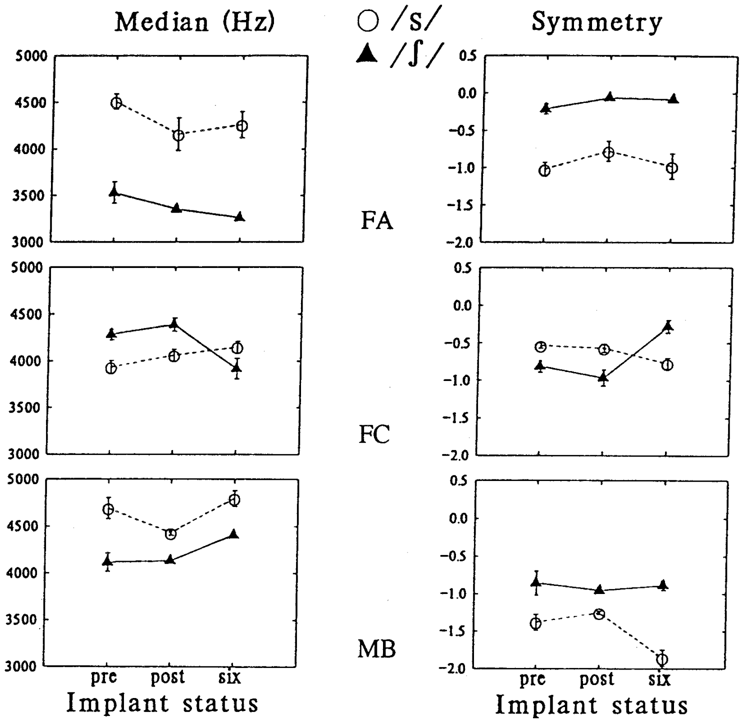

Figure 6. Spectral median (left panels) and symmetry (right panels) for one male (MB) and two female (FA, FC) CI users.

3.3 There are systematic relations among changes in perception, production and intelligibility.

We have hypothesized that if changes are observed in phonemic contrasts after speakers gain hearing with a CI, those changes are in the direction of improved contrast and they are driven by the speaker's motivation to enhance intelligibility. To test this hypothesis in some detail, we gathered speech production, perception and intelligibility data for the liquids /r/ and /l/ spoken in carrier phrases by eight postlingually deaf adults, pre- and post implant with a CI. Formant transition analysis for the CI speakers and two speakers with normal hearing indicated that /r/ and /l/ could be differentiated by the extent of the F3 transition from vowel beginning to mid vowel and the distance in Hz between F2 and F3 at the C-V boundary. Speakers who had a limited contrast between /r/ and /l/ pre-implant and who showed improvement in their perception of these consonants with prosthetic hearing were found to demonstrate greatly improved production of /r/ and /l/ six months post-CI. The speech production changes noted in the acoustic analyses were corroborated by intelligibility improvements in the post-CI speech, as measured with a panel of normal-hearing listeners. Figure 7 shows an example of enhanced contrast between /r/ and /l/ for a male CI user, from pre- to post implant (Matthies et al., submitted). We have also observed similar results for vowels: significant covariation of perception and production, and of production and intelligibility (Vick et al., 2001a; also see Gould et al., 2001).

Intelligibility is usually quite good in postlingually deafened candidates for cochlear implants; nevertheless, we have observed gains in sentence intelligibility in subjects who had some room for improvement (Vick et al., 2001b). Such findings indicate that speakers are quite sensitive to intelligibility decrements after implantation and presumably can retune neural mappings even with the relatively crude auditory stimulation supplied by implants.

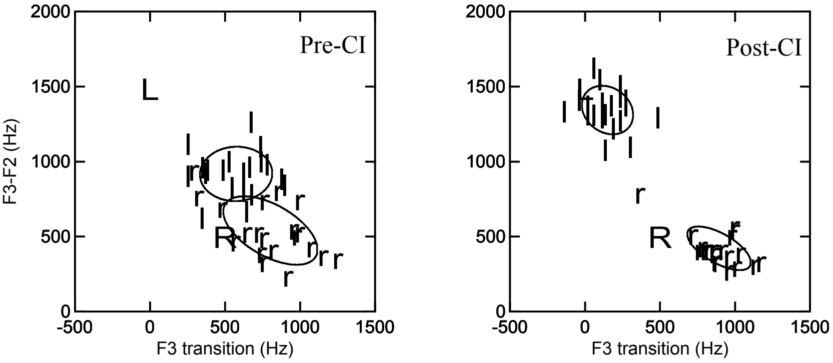

Figure 7. Separation between F3 and F2 at vowel beginning vs. extent of F3 transition from vowel beginning to mid-vowel for preceding /r/ and /l/ in a carrier phrase, spoken by a male CI user. (Capital letters indicate normative data.) Left half – pre-implant; right half – data pooled from 6 and 12 months post.

3.4 A change in the vocal tract may invalidate systemic mappings.

We have made another observation of the /s-ʃ/ contrast, from a subject, FD, who lost hearing due to bilateral acoustic neuromas (NF-2). The subject had tumor-removal surgery that severed her remaining auditory nerve (Perkell et al., 1995a). At the time of surgery, she received an early version of an auditory brainstem implant, which effectively provided her with auditory envelope but not spectral cues. Figure 8 shows spectral median versus week from her onset of hearing loss (OHL) for /s/ and /ʃ/. Before OHL and continuing for over 70 weeks post-OHL, FD maintained a good contrast between the two sounds.

At week 72, FD had another surgery, this time to anastomose her left hypoglossal nerve to the facial nerve, in an attempt to restore some facial function that had also been lost at the time of tumor removal surgery. The anastomosis surgery denervated some tongue muscles on the left side, producing a slight tongue weakness that effectively altered a functional property of the vocal tract. Without auditory feedback about the sibilant contrast to help the control mechanism develop a compensatory adaptation to the tongue weakness, the contrast gradually collapsed. In terms of the model in Figure 1, the anastomosis surgery invalidated systemic mappings between auditory and articulatory parameters by changing a characteristic of the low-level control of the “biomechanical plant”. Due to the subject’s hearing loss, it was then impossible for her to update the mappings by making auditorily based adjustments. Since people with normal hearing are capable of compensating for significant changes in vocal-tract morphology (e.g. with the initial insertion of dentures), we assume that if this speaker had adequate hearing, she would have been able to compensate for the surgery, even though it resulted in some slight tongue weakness.

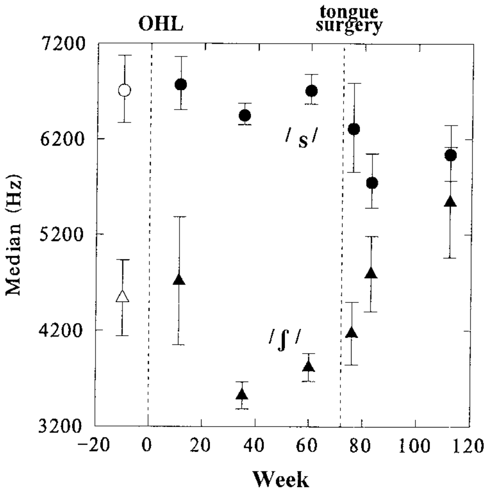

Figure 8. Spectral median for /s/ and /ʃ/ versus time in weeks from a subject who lost hearing (at time OHL) due to removal of an acoustic neuroma.

The fact that the collapse in contrast was gradual after the tongue surgery may be due to the speaker relying on a combination of auditory and somatosensory feedback. Without auditory spectral feedback to reinforce the somatosensory input, the latter gradually became inadequate to maintain the contrast. Although the model as presented in this paper deals primarily with auditory goals and feedback, other versions incorporate the idea that the goals for some sounds, especially consonants, include orosensory components (Guenther, 1995a; Perkell, 1997).

3.5 Language-Specific, Hearing-Related Changes in Vowel Spaces may reflect a tradeoff between phonemic contrast and economy of effort.

Another important feature of the model summarized in Figure 1 is that movements for a sequence of phonemes are planned to afford an economy of effort (Guenther, 1995a; also see Lindblom, 1983; Keating, 1990). This is accomplished by planning an auditory trajectory that passes through the parts of the auditory goal regions that are closest to those of the neighboring sounds in the sequence. In this study, two hypotheses were tested that are derived from the view that vowel production is influenced by competing demands of intelligibility for the listener and least effort in the speaker: (1) Hearing enables a CI user to produce vowels distinctly from one another; without hearing, the speaker may give more weight to economy of effort, leading to reduced vowel distinctiveness. (2) Speakers may need to produce vowels more distinctly from one another in a language with a relatively “crowded” vowel space, such as American English, than in a language with relatively few vowels, such as Spanish. Thus, when switching between hearing and non-hearing states, English speakers may show a tradeoff between vowel distinctiveness and least effort, while Spanish speakers may not. To test these hypotheses, we predicted that there would be a reduction of average vowel spacing (AVS – average inter-vowel distance in the F1-F2 plane) with interrupted hearing for English-speaking CI users, but no systematic change in AVS for Spanish CI users. We recorded vowel productions of seven English- and seven Spanish-speaking CI users, who had been using their implants for at least one year. When their implant speech processors were turned off and on several times in two sessions, we found that AVS was consistently larger for the English speakers with hearing than without hearing. The presence and direction of AVS change was more variable for the Spanish speakers, both within and between subjects. Thus, vowel distinctiveness was enhanced with the provision of some hearing in the language group with a more crowded vowel space but not in the language group with fewer vowels. This result supports the view that speakers seek to minimize effort while maintaining the distinctiveness of auditory goals.

3.6 Changes in phonemic contrasts can be quite rapid.

We have observed previously that the kinds of contrast changes cited above can occur quite rapidly, almost as soon as a CI user’s speech processor is switched on or off (Svirsky, et al., 1992). In this study, we investigated such changes in more detail, comparing changes in vowel SPL and duration with those in phonemic contrasts between the vowels /ɛ/ and /æ/ and the sibilants /s/ and /ʃ/. An apparatus was built to switch the speech processor of a subject’s implant on and off a number of times in a single experimental session while the subject repeated a large number of utterances containing the target sounds. Two normal-hearing subjects performed the same paradigm, except that the “on” condition consisted of hearing their own speech fed back to them over a set of headphones and the “off” condition consisted of hearing loud noise that masked their speech. Using the times of on-off or off-on switches as line-up points for averaging, parameters were compared across the switches.

The speakers’ vowel SPL and duration had changed by the first utterance following the switch. The top half of Figure 9 shows changes in vowel contrast, defined as the Euclidian distance in F1-F2 space between /ɛ/ and /æ/. The bottom half shows change in sibilant contrast, defined as the difference between the spectral median of /s/ and /ʃ/. There are four bars for each subject, corresponding to the change from the off to on condition averaged over 50 seconds on either side of the switch (1), the change from the tokens immediately before to immediately after the switch (2), and the corresponding on-to-off changes (3 and 4). Shaded bars represent significant changes ($p < .05$). All the significant changes but two (vowel contrasts for subject FH, bars 1 and 3) are consistent with the prediction that contrast is enhanced in the presence of hearing and diminished without hearing. Further, the changes in phonemic parameters are just as rapid as those for SPL. The model has a mechanism for making phonemic changes as rapid as those described here. This is done using a control parameter that rapidly scales the sizes of all auditory goal regions (Guenther, 1995a).

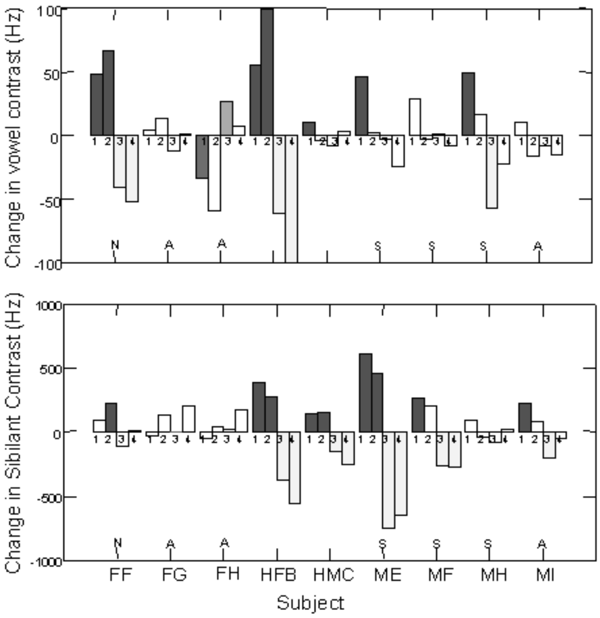

Figure 9. Top half: Change in vowel contrast, defined as the Euclidian distance in F1-F2 space between /ɛ/ and /æ/. Bottom half: Change in sibilant contrast, defined as the difference between the spectral median of /s/ and /ʃ/. Subjects HFB and HMC have normal hearing. The darkest gray bars indicate significant off-to-on changes in the predicted direction; the lightest gray bars indicate significant on-to-off changes; the bars with intermediate shading indicate significant changes opposite to the predicted direction.

4. Summary and Discussion

Section 3 described studies and interpretations that are largely compatible with the model of speech motor control outlined in Sections 1 and 2. In the model, speech movements are planned to achieve sequences of auditory goals, in a way that provides an economy of effort. The planning utilizes two kinds of mappings, phoneme-specific (language-specific) and systemic (speaker-specific). The mappings are acquired with the use of auditory feedback and rely on auditory feedback for their maintenance over the course of a lifetime. In general, the experimental results support the interpretation that, once learned, the neural mappings are stable, even in the absence of hearing. In the collapse of the sibilant contrast in an NF2 patient who had lost access to auditory spectral information, we suggested that the collapse was due to a change in the vocal tract that could not be compensated for without the needed auditory feedback. The finding of systematic increases in average vowel spacing with the provision of hearing by a CI in English speakers, but not in Spanish speakers, may reflect a tradeoff between phonemic contrast and economy of effort, both of which are predicted by the model to guide articulatory trajectories. Studies of vowels and the /r/-/l/ contrast demonstrated systematic relations among changes in perception, production and intelligibility, leading to the inference that speakers are quite sensitive to decrements in their intelligibility and can correct them with access to even the relatively crude auditory stimulation provided by a CI. Finally, the rapid changes in phonemic contrasts described in Section 3.6 suggest that, although full retuning of systemic mappings may take a long time, some aspects of the phoneme-specific mapping can be changed very rapidly. In the model, changes this abrupt can be implemented by a single neural control signal that effectively shrinks or expands all the goal regions simultaneously.

We are still far from being able to establish quantitative relations between the modeling and experimental work in this area. Nevertheless, when considered collectively, our experimental results provide strong qualitative support for the modeling approach, and as exemplified by the above-cited examples, they can provide insights that may strongly influence model development. In the long run, we believe that this combination of modeling and experimentation will move us closer to understanding neural processes underlying relations between speech perception and production, and to a coherent account of speech processes in people with normal speech and hearing and in clinical populations.

References

Bailly, G., Laboissière, R., and Schwartz, J.L. (1991). Formant trajectories as audible gestures: An alternative for speech synthesis. *Journal of Phonetics*, **19**, pp. 9-23.

Bookheimer, S.Y., Zeffiro, T.A., Blaxton, T., Gaillard, W., and Theodore, W. (1995). Regional cerebral blood flow during object naming and word reading. *Human Brain Mapping*, **3**, pp. 93-106.

Callan, D., Kent, R., Guenther, F. H., and 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.

Caplan, D., Gow, D., and Makris, N. (1995). Analysis of lesions by MRI in stroke patients with acoustic-phonetic processing deficits. *Neurology*, **45**, pp. 293-298.

Celsis, P., Boulanouar, K., Ranjeva, J.P., Berry, I., Nespoulous, J.L., and Chollet, F. (1999). Differential fMRI responses in the left posterior superior temporal gyrus and left supramarginal gyrus to habituation and change detection in syllables and tones. *NeuroImage*, **9**, pp. 135-144.

Corfield, D.R., Murphy, K., Josephs, O., Fink, G.R., Frackowiak, R.S.J., Guz, A., Adams, L., and Turner, R. (1999). Cortical and subcortical control of tongue movement in humans: A functional neuroimaging study using fMRI. *Journal of Applied Physiology*, **85**, pp. 1468-1477.

Damasio, H., and Damasio, A.R. (1980). The anatomical basis of conduction aphasia. *Brain*, **103**, pp. 337-350.

Daniloff, R., Schuckers, G., and Feth, L. (1980). *The physiology of speech and hearing: An introduction*. Englewood Cliffs NJ: Prentice-Hall.

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

Dronkers, N.F. (1996). A new brain region for coordinating speech articulation. *Nature*, **384**, pp. 159-161.

Fox, P.T., Huang, A., Parsons, L.M., Xiong, J., Zamarippa, F., Rainey, L., and Lancaster, J.L. (2001). Location-probability profiles for the mouth region of human primary motor-sensory cortex: Model and validation. *NeuroImage*, **13**, pp. 196-209.

Geschwind, N. (1965). Disconnexion syndromes in animals and man. I. *Brain*, **88**, pp. 237-294.

Gould, J., Lane, H., Perkell, J., Vick, J., Matthies, M., and Zandipour, M. (2001). Changes in the intelligibility of postlingually deaf adults after cochlear implantation. *Ear and Hearing*, **22**, pp. 453-60.

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

Guenther, F.H. (1995a). Speech sound acquisition, coarticulation, and rate effects in a neural network model of speech production. *Psychological Review*, **102**, pp. 594-621.

Guenther, F.H. (1995b). A modeling framework for speech motor development and kinematic articulator control. *Proceedings of the XIIIth International Conference of Phonetic Sciences (*vol. 2, pp. 92-99). Stockholm, Sweden: KTH and Stockholm University.

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

Guenther, F.H., and Micci Barreca, D. (1997). Neural models for flexible control of redundant systems. In: P. Morasso and V. Sanguineti (eds.), *Self-organization, Computational Maps, and Motor Control* (pp. 383-421). Amsterdam: Elsevier-North Holland.

Hickok, G., Erhard, P., Kassubek, J., Helms-Tillery, A.K., Naeve-Velguth, S., Strupp, J.P., Strick, P.L., and Ugurbil, K. (2000). A functional magnetic resonance imaging study of the role of left posterior superior temporal gyrus in speech production: Implications for the explanation of conduction aphasia. *Neuroscience Letters*, **287**, pp. 156-160.

Keating, P.A. (1990). The window model of coarticulation: Articulatory evidence. In J. Kingston and M. E. Beckman (Eds*.), Papers in laboratory phonology I: Between the grammar and physics of speech* (pp. 451-470). Cambridge: Cambridge University Press.

Kent, R. D. and Minifie, F. D. (1977). Coarticulation in recent speech production models. *Journal of Phonetics*, **5**, pp. 115-133.

Levelt, W.J.M., Praamstra, P., Meyer, A.S., Helenius, P., and Salmelin, R. (1998). An MEG study of picture naming. *Journal of Cognitive Neuroscience*, **10**, pp. 553-567.

Liégeois, A. (1977). Automatic supervisory control of the configuration and behavior of multibody mechanisms. *IEEE Transactions on Systems, Man, and Cybernetics,* **7**(12), 869-871.

Lindblom, B. (1983). Economy of speech gestures. In P. F. MacNeilage (Ed.), *The production of speech* (pp. 217-245). New York: Springer-Verlag.

Lindblom, B., and MacNeilage, P. F. (1986). Action theory: Problems and alternative approaches. *Journal of Phonetics*, **14**, pp. 117-132.

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

Matthies, M.L., Svirsky, M.A., Lane, H., and Perkell, J.S. (1994). A preliminary study of the effects of cochlear implants on the production of sibilants, *Journal of the Acoustical Society of America*, **96**, pp. 1367-1373.

Matthies, M., Vick, J., Perkell, J., Lane, H., Zandipour, M., and Gould, J. (2003). Effects of cochlear implants on the speech production, perception, and intelligibility of the liquids /r/ and /l/, in preparation.

Paus, T., Perry, D.W., Zatorre, R.J., Worsley, K.J., and Evans, A.C. (1996). Modulation of cerebral blood flow in the human auditory cortex during speech: Role of motor-to-sensory discharges. *European Journal of Neuroscience*, **8**, pp. 2236-2246.

Perkell, J.S. (1997). Articulatory Processes, in W. Hardcastle and J. Laver (Eds.) *The Handbook of Phonetic Sciences*, (pp. 333-370). Oxford, UK: Blackwell.

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

Perkell, J., Lane, H., Svirsky, M. and Webster, J. (1992). Speech of cochlear implant patients: A longitudinal study of vowel production, *Journal of the Acoustical Society of America,* **91**, pp. 2961-2979.

Perkell, J., Manzella, J., Wozniak, J., Matthies, M., Lane, H. Svirsky, M. Guiod, P. Delhorne, L. Short, P. MacCollin, M. and Mitchell, C. (1995a). Changes in speech production following hearing loss due to bilateral acoustic neuromas, *Proceedings of the XIIIth International Congress of Phonetic Sciences*, vol. 3, (pp. 194-197). Stockholm.

Perkell, J. S., Matthies, M. L., Svirsky, M. A., and Jordan, M. I. (1995b). Goal-based speech motor control: A theoretical framework and some preliminary data. *Journal of Phonetics,* **23**, pp. 23-35.

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

Peterson, G. E. and Barney H. L. (1952) Control methods used in a study of the vowels, *Journal of the Acoustical Society of America*, **24**, pp. 175-184.

Picheny, M. A., Durlach, N. I., and Braida, L. D. (1985). Speaking clearly for the hard of hearing I: Intelligibility differences between clear and conversational speech. *Journal of Speech and Hearing Research*, **28**, pp. 96-103.

Picheny, M. A., Durlach, Nathaniel .I., and Braida, L. D. (1986). Speaking clearly for the hard of hearing II: Acoustic characteristics of clear and conversational speech. *Journal of Speech and Hearing Research*, **29**, pp. 434-446.

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

Svirsky, M., Lane, H., Perkell, J., and Webster, J. (1992). Speech of cochlear implant patients: Results of a short-term auditory deprivation study. *Journal of the Acoustical Society of America,* **92**, pp. 1284-1300.

Vick, J., Lane, H., Perkell, J., Matthies, M., Gould, J., and Zandipour, M. (2001a). Speech perception, production and intelligibility improvements in vowel-pair contrasts in adults who receive cochlear implants. *Journal of Speech, Language and Hearing Research*, **44**, pp. 1257-68.

Vick, J., Lane, H., Perkell, J., Zandipour, M. and Matthies, M. (2001b). Sentence Intelligibility in adults who receive cochlear implants. Poster presented at the 2001 Conference on Implantable Auditory Prostheses, Asilomar, CA.

Wise, R.J., Greene, J., Buchel, C., and Scott, S.K. (1999). Brain regions involved in articulation. *Lancet*, **353**, pp. 1057-1061.

Acknowledgements

This research was supported by the National Institute on Deafness and other Communication Disorders (grants R01 DC02852 to F. Guenther; R01 DC03007 to J. Perkell).