1997-02

A Theoretical Investigation of Reference Frames for the Planning of Speech Movements

A Theoretical Investigation of Reference Frames for the Planning of Speech Movements

Running title: Speech reference frames

Frank Guenther, Michelle Hampson, and Dave Johnson

February 1997

Technical Report CAS/CNS-97-002

Boston University Center for Adaptive Systems and
Department of Cognitive and Neural Systems
677 Beacon Street
Boston, MA 02215

A Theoretical Investigation of Reference Frames for the Planning of Speech Movements

Running title: Speech reference frames

Frank H. Guenther*
Michelle Hampson
Dave Johnson

Boston University
Center for Adaptive Systems and
Department of Cognitive and Neural Systems
677 Beacon Street
Boston, MA, 02215
Fax Number: (617) 353-7755
Email: guenther@cns.bu.edu

Submitted to *Psychological Review*

*Frank Guenther supported in part by the Alfred P. Sloan Foundation and the National Institutes of Health (1 R29 DC02852-01). Michelle Hampson supported in part by the National Institutes of Health (1 R29 DC02852-01). Dave Johnson supported in part by the Office of Naval Research (ONR N00014-94-1-0597, ONR N00014-95-1-0409, and ONR N00014-95-1-0657). The authors would like to thank Seth Cameron, Daniele Micci Barreca, Pascal Perrier, and Joe Perkell for their comments and contributions to this manuscript.

ABSTRACT

Does the speech motor control system utilize invariant vocal tract shape targets of any kind when producing phonemes? We present a four-part theoretical treatment favoring models whose only invariant targets are auditory perceptual targets over models that posit invariant constriction targets. When combined with earlier theoretical and experimental results (Guenther, 1995a,b; Perkell et al., 1993; Savariaux et al., 1995a,b), our hypothesis is that, for vowels and semi-vowels at least, the only invariant targets of the speech production process are multidimensional regions in auditory perceptual space. These auditory perceptual target regions are hypothesized to arise during development as an emergent property of neural map formation in the auditory system (Guenther and Gjaja, 1996), as evidenced by the perceptual magnet effect. Furthermore, speech movements are planned as trajectories in auditory perceptual space. These trajectories are then mapped into articulator movements through a neural mapping that allows motor equivalent variability in constriction locations and degrees when needed, but maintains approximate constriction invariance for a given sound in most instances. These hypotheses are illustrated and substantiated using computer simulations of the DIVA model of speech acquisition and production. Finally, we pose several difficult challenges to proponents of constriction theories based on this theoretical treatment.

1. Introduction: Reference frames and the targets of speech production

Does the speech motor control system utilize invariant vocal tract shape targets of any kind when producing phonemes? In other words, when moving the articulators to produce a speech sound, is the sole invariant goal of the movements an auditory goal, or does the brain effectively equate the auditory goal with invariant aspects of vocal tract shape that serve as the target of movement? Relatedly, are the articulator movement trajectories planned within an auditory reference frame, or are they planned within some reference frame more closely related to the articulators or the shape of the vocal tract?

The answers to these questions have important implications for research in motor control, speech, and phonology. First, they are at the heart of an important debate in motor control. In addition to the treatment of these questions in the speech motor literature (e.g., Bailly, Laboissière, and Schwartz, 1991; Browman and Goldstein, 1990a,b; Guenther, 1995a,b; Laboissière and Galvan, 1995; Perkell, 1997; Perkell, Matthies, Svirsky, and Jordan, 1993; Saltzman and Munhall, 1989; Savariaux, Perrier, and Orliaguet, 1995a), analogous questions have frequently arisen in the research literature on reaching movements. The analog of the first question posed above in the domain of reaching movements can be stated as follows: is the sole invariant target of a reaching movement the spatial position of the hand at the end of the reach, or does the motor system utilize an arm configuration or posture target (e.g., a set of joint angles)? An extreme version of the former view is found in the DIRECT model (Bullock, Grossberg, and Guenther, 1993; Guenther, 1992), an extreme version of the latter view is found in the KNOWLEDGE model (Rosenbaum, Engelbrecht, Bushe, and Loukopoulos, 1993; Rosenbaum et al., 1995), and supporting evidence for each view has been presented in these works[1]. The arm movement analog of the second question posed above is: are reaching movement trajectories planned in a spatial reference frame, or are they planned in a reference frame more closely related to joint angles or muscle lengths? Several investigators have provided data suggesting that movement trajectories are planned in a spatial reference frame rather than a reference frame more closely related to the joints or muscles (e.g., Morasso, 1981; Wolpert, Ghahramani, and Jordan, 1994, 1995) and other researchers have provided computational models based on spatial trajectory planning to account for these data (e.g., Bullock, Grossberg, and Guenther, 1993; Flash, 1989; Guenther, 1992; Hogan, 1984). Other theorists have suggested that trajectories are planned in a reference frame more closely related to the joints (e.g., Cruse, 1986; Rosenbaum, et al., 1993, 1995; Uno, Kawato, and Suzuki, 1989) and have provided experimental evidence that appears to contradict some of the spatial trajectory planning proposals (e.g., Gomi and Kawato, 1996). Still other theories suggest some combination of spatial planning and joint/muscle influences (Cruse, Brüwer, and Dean, 1993; Guenther and

1. See Guenther and Micci Barreca (1997) for a more thorough treatment of existing data on postural targets and their implications for reaching models.

Micci Barreca, 1997). For example, Guenther and Micci Barreca (1997) suggest that the only invariant target for a reaching movement is a spatial target of the hand and that movement trajectories are planned in spatial coordinates, but the mapping from planned spatial trajectories to the muscle contractions needed to carry them out contains biases that favor certain arm postures over others. The model we propose in the current article can be thought of as a speech production analog of this proposal, as detailed in Section 3.

Second, an understanding of the reference frame for speech motor planning is crucial for efficient acquisition and analysis of speech production data. A primary goal of speech research is to build a mechanistic understanding of the neural processes underlying speech perception, speech production, and the interactions between perception and production. The clearer our understanding of the reference frame used to plan speech movements, the easier it is to design useful experiments and interpret the resulting data. For example, the large amount of articulatory variability seen for the American English phoneme /r/ has long been a source of difficulty for speech researchers. The conviction that phoneme production utilizes vocal tract shape targets of some sort, coupled with the fact that the same speaker often uses two completely different shapes to produce this sound in different phonetic contexts (Delattre and Freeman, 1968; Espy-Wilson and Boyce, 1994; Hagiwara, 1994, 1995; Narayanan, Alwan, and Haker, 1995; Ong and Stone, 1997; Westbury, Hashi, and Lindstrom, 1995), has led several researchers to attempt to characterize the different /r/ productions as different classes of /r/. Interestingly, this has led to several different answers to the question of just how many classes of /r/ there are. Delattre and Freeman (1968) break /r/ productions into eight different classes while suggesting a sort of continuum of /r/ productions, Espy-Wilson and Boyce (1994) and Ong and Stone (1997) interpret their data in terms of two variants of /r/, whereas Hagiwara (1994) suggests that three variants of /r/ exist. In the current paper, however, we show how /r/ variability can be explained much more simply if it is assumed that the reference frame for speech movement planning is auditory; i.e., the only invariant target of the production process for /r/ is an auditory target, not a vocal tract shape target. Embedding this idea into the DIVA model of speech production (Guenther, 1994, 1995a,b) leads to a simple explanation in which a single invariant target for /r/ results in different /r/ articulations depending on the shape of the vocal tract when /r/ production commences; i.e., depending on phonetic context. This explanation also accounts for the difficulty in determining the number of classes of /r/ in previous studies. These topics are treated in detail in Section 5.

Third, the questions posed in the introductory paragraph have important implications for the well known "motor theory" of speech perception (e.g., Liberman, Cooper, Shankweiler, and Studdert-Kennedy, 1967; Liberman and Mattingly, 1985). The motor theory states that invariant articulatory gestures or motor commands underly the perception of speech. In other words, the speech perception system effectively consults with the gestural targets of the production system when identifying speech sounds. The motor theory has been attacked over the years from several different angles (see Lieberman and Blumstein, 1988). If it turns out that

even the speech *production* process utilizes no invariant articulatory or vocal tract constriction targets, but instead uses only targets that are more directly related to the acoustic signal as suggested herein (at least for vowels and semivowels; see also Bailly et al., 1991; Perkell et al., 1993; Savariaux, Perrier, and Orliaguet, 1995a), then the motor theory claim that the speech *perception* system utilizes an invariant articulatory gesture representation rests on even shakier ground.

Finally, the answers to these questions are important from the viewpoint of at least one major phonological theory, the "articulatory phonology" of Browman and Goldstein (e.g., 1990a,b). This theory posits that the basic units of phonetics and phonology are dynamically-defined articulatory gestures. In their linguistic gestural model, they further define the reference frame for these gestures to be a vocal tract constriction reference frame, and the invariant targets of speech production are characterized as vocal tract constriction targets rather than acoustic/auditory targets. The linguistic gestural model, in concert with the task dynamic model of Saltzman and Munhall (1989), has served as the most complete and influential description of the speech production process over the past several years. The question of whether the phonetic units and invariant targets of speech production are better characterized as constriction gestures or as acoustic/auditory targets is still an open one, however, and we suggest herein that, for some sounds at least, the invariant targets are better characterized as auditory perceptual targets, not constriction targets.

This article provides a theoretical treatment of the questions posed in the introductory paragraph based on a wide range of speech production data. This treatment stems from the viewpoint that the simplest explanation for the range of existing speech production data is that the speech motor control system utilizes invariant auditory perceptual targets when producing phonemes, and that movement trajectories to these targets are planned in an auditory perceptual space. It is further suggested that apparent invariances in constriction location and degree may well arise due to biases in the mapping from planned auditory perceptual trajectories to the muscle contractions that carry them out, rather than as the result of any invariant constriction targets. Computer simulations of the DIVA model of speech production (Guenther, 1994, 1995a,b) are used to illustrate and substantiate these claims. The treatment herein concentrates on vowels and semivowels; the situation for consonants is less clear at present and will only be addressed briefly in the concluding remarks.

Before proceeding with the theoretical discussion, it is useful to more precisely define the different reference frames considered in this article. This will be done with reference to the block diagram of the DIVA model shown in Figure 1. The DIVA model is a neural network architecture that learns various mappings (shown as filled semicircles in Figure 1) between reference frames during a babbling cycle. After babbling, the model is capable of producing arbitrary combinations of the phonemes it has learned during babbling. The implementation described herein produces only vowels and /r/. The following paragraphs will concentrate on the aspects

of the model that are relevant to the current article; see Guenther (1995a) for a more complete description of the model's learning processes and properties.

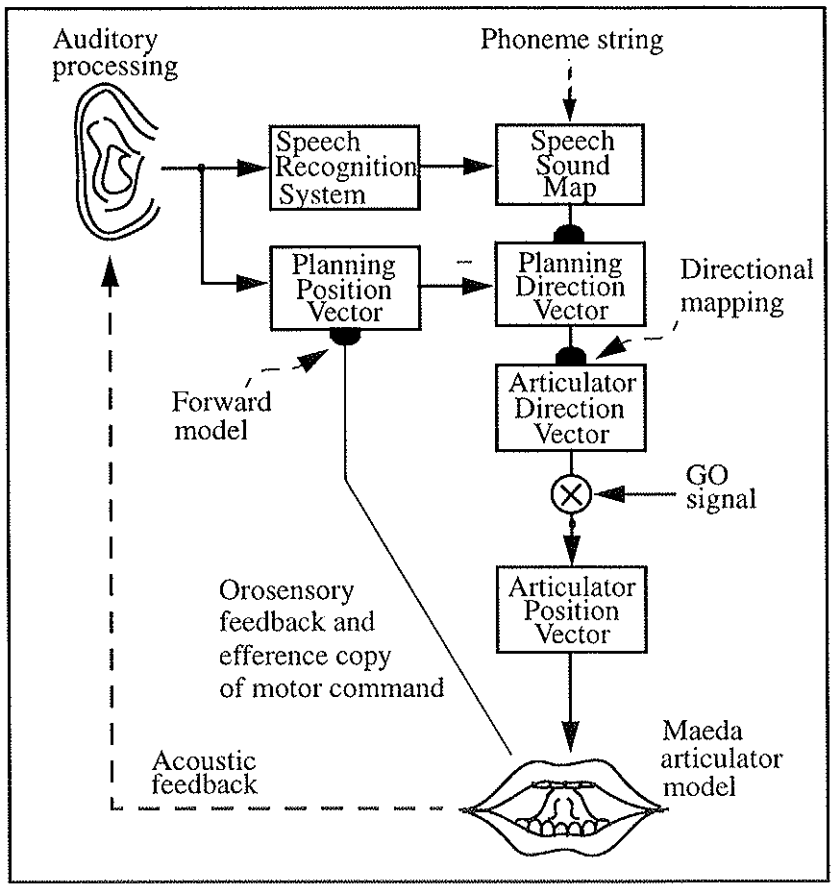

FIGURE 1. Overview of the DIVA model of speech acquisition and production. Neural mappings learned during babbling are indicated by filled semicircles. See text for details.

Muscle length reference frame. This frame of reference describes the lengths and shortening velocities of the muscles that move the speech articulators. At some level of the motor control process, muscle lengths or contractile states must be coded in the nervous system in order to position the speech articulators. However, this does not imply that the speech motor system utilizes an *invariant* muscle length target for each speech sound, and in fact much experimental data speak against this kind of target. For example, insertion of a bite block between the teeth forces a completely different set of muscle lengths to produce the same vowel sound, yet people are capable of compensating for bite blocks even on the first glottal pulse (Lindblom, Lubker, and Gay, 1979), illustrating the motor system's capacity to use different muscle length configurations to produce the same phoneme under different conditions.

Articulator reference frame. An articulator reference frame, or articulator space, refers to a reference frame whose coordinates roughly correspond to the primary movement degrees of freedom of the speech articulators (e.g., Mermelstein, 1973; Rubin, Baer, and Mermelstein, 1981; Maeda, 1990). Although it is clear that the primary movement degrees of freedom are closely related to the musculature, the articulator reference frame is often assumed to be of lower dimensionality than the muscle reference frame. For example, several muscles may move together in a synergy that corresponds to a single movement degree of freedom. For the purposes of this article, the distinction between an articulator reference frame and a muscle length reference frame is relatively unimportant, and we will therefore typically equate the two. The distinction becomes more important, however, for lower-level modeling of the kinematics and dynamics of the speech articulators (e.g., Laboissière, Ostry, and Perrier, 1995; Ostry, Gribble, and Gracco, 1996; Stone, 1991; Wilhelms-Tricarico, 1995, 1996).

The Articulator Direction Vector and Articulator Position Vector in Figure 1 act as commands that move the speech articulators in the model. These vectors each have seven dimensions, corresponding to the seven degrees of freedom (DOFs) of the Maeda articulator model (Maeda, 1990), which has been embedded in the DIVA framework. The DOFs are for the jaw (1 DOF), the tongue (3 DOFs), the lips (2 DOFs) and the larynx height (1 DOF). The positions of the articulators are used to synthesize an acoustic signal using the Maeda model. Movement trajectories planned by the DIVA model in auditory space (discussed below) are mapped into movement directions of the articulators at the Articulator Direction Vector stage. These directional commands are then used to update the articulator positions at the Articulator Position Vector stage. The GO signal (Bullock and Grossberg, 1988) in Figure 1 controls movement speed by determining how quickly the articulators are moved in the direction specified by the Articulator Direction Vector; see Guenther (1995a) for details.

Tactile reference frame. This reference frame describes the states of pressure receptors on the surfaces of the speech articulators. For example, the pressure produced when the tongue tip is pressed against the hard palate is registered by neural mechanoreceptors in the tongue and the palatal surface. Mechanoreceptors provide important information about articulator positions when contact between articulators is made, but provide little or no information when contact is absent. No tactile information is used in the current implementation of the DIVA model, largely because the model is not being used to produce consonants. Previous versions of the model have used tactile information from a more simplistic articulator set (Guenther, 1994, 1995a). We will occasionally use the term "orosensory information" (c.f. Perkell, 1980) to refer to a combination of tactile and muscle length information.

Constriction reference frame. Several researchers have proposed reference frames for speech production whose coordinates describe the locations and degrees of key constrictions in the vocal tract (e.g., Browman and Goldstein, 1990a,b; Coker, 1976; Guenther, 1994, 1995a; Kroger, 1993; Saltzman and Mun-

hall, 1989). Typical constrictions include a tongue body constriction, tongue tip constriction, and lip constriction. It is important to note that the relationship between the constriction frame and the articulator frame is one-to-many; that is, a given set of constriction locations and degrees can be reached by an infinite number of different articulator configurations. 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. This one-to-many relationship makes it possible for a movement controller that uses invariant constriction targets and an appropriate mapping between the constriction and articulator frames to overcome constraints on the articulators (such as a bite block) by utilizing a different articulator configuration than usual to produce the same constrictions as usual (e.g., Guenther, 1992, 1994, 1995a; Saltzman and Munhall, 1989). This ability to use different movements to reach the same goal under different conditions, called *motor equivalence*, is a ubiquitous property of biological motor systems and is addressed further in Section 4.

Acoustic reference frame. The acoustic reference frame describes the properties of the acoustic signal produced by the vocal tract (e.g., formant frequencies, amplitudes, and bandwidths). Strictly speaking, the central nervous system has access to the acoustic signal only after transduction by the auditory system. However, several researchers have used the word "acoustic" to refer to this transduced signal (e.g., Guenther, 1995b; Perkell et al., 1993). In the current paper we will use the more precise term "auditory perceptual" to refer to the transduced version of the acoustic signal (c.f. Miller, 1989; Savariaux, Perrier, & Schwartz, 1995b).

Auditory perceptual reference frame. In the block diagram of Figure 1, the acoustic signal is transduced into an auditory perceptual reference frame by the auditory system, and the resulting auditory perceptual information projects to a speech recognition system that identifies speech sounds. Although the important aspects of the auditory representation for speech perception and production are still not fully understood, several researchers have attempted to characterize them. In the current implementation of the DIVA model, we utilize the auditory perceptual frame proposed by Miller (1989), although we acknowledge the incompleteness of this auditory representation for capturing all of the perceptually important aspects of speech sounds. This auditory perceptual space is made up of three dimensions x_i:

$$x_1 = \log\left(\frac{F1}{SR}\right) \tag{1}$$

$$x_2 = \log\left(\frac{F2}{F1}\right) \tag{2}$$

$$x_3 = \log\left(\frac{F3}{F2}\right) \tag{3}$$

where F1, F2, and F3 are the first three formants of the acoustic signal, and $SR = 168(F0/168)^{1/3}$, where F0 is the fundamental frequency of the speech waveform. This space was chosen by Miller in part due to the fact that these coordinates remain relatively constant for the same vowel when spoken by men, women, and children, unlike formant frequencies.

We also hypothesize that the auditory perceptual reference frame is used to plan speech movement trajectories, as indicated by the arrow to the Planning Position Vector stage in Figure 1. This replaces the constriction-based planning frame used in earlier versions of the DIVA model (Guenther, 1994, 1995a). The Planning Position Vector in the model represents the current state of the vocal tract within the auditory perceptual reference frame. This can be determined from acoustic feedback or from the output of a "forward model" (c.f. Jordan, 1990) that transforms orosensory feedback and/or an efference copy of the articulator position commands into the auditory perceptual reference frame. (See Section 2 for further discussion of the forward model concept.) Projections from the Speech Sound Map to the Planning Direction Vector stage encode a learned auditory perceptual target for each sound. These targets take the form of multidimensional regions, rather than points, in auditory perceptual space (see also Perkell et al., 1997). Guenther (1995a) shows how a region theory for the targets of speech provides a unified explanation for a wide range of speech production phenomena, including data on motor equivalence, speaking rate effects, carryover coarticulation, and anticipatory coarticulation. Guenther and Gjaja (1996) hypothesize that these auditory perceptual target regions arise during development as an emergent property of neural map formation in the auditory system, as evidenced by the perceptual magnet effect (Kuhl, 1991, 1995; Iverson and Kuhl, 1995).

The current state of the vocal tract is compared to the auditory perceptual target region at the Planning Direction Vector stage. The cell activities at the Planning Direction Vector stage represent the desired movement direction in auditory perceptual coordinates (i.e., the movement direction needed to get to the nearest point on the target region). The time course of these activities represents the planned movement trajectory in auditory perceptual coordinates, and this trajectory is then transformed into appropriate movements of the speech articulators through the learned mapping projecting from the Planning Direction Vector to the Articulator Direction Vector.

This directional mapping from the auditory perceptual frame to the articulator frame is a key component of the DIVA model. Note that the model maps desired movement *directions* in auditory perceptual space into movement directions of the articulators, rather than mapping target *positions* in auditory perceptual space into articulator configurations. Because of this, the model does not have a fixed articulator configuration for each position in auditory perceptual space. Instead, it can use many different articulator configurations (infinitely many, in fact) to reach a given position in auditory perceptual space. (Like the relationship between constrictions and articulator configurations, the relationship between points in auditory perceptual space and articulator configurations is one-to-many.) In short, the

use of a directional mapping leads to the property that *the only invariant target for a speech sound is the auditory perceptual target*, and this target can be reached with an infinite number of different articulator configurations or vocal tract constriction configurations depending on things like phonetic context or constraints on the articulators. This point is central to much of the discussion in the remainder of this article.

The primary contention of this article is that, although the idea of invariant vocal tract constriction targets has led to a much better understanding of speech production over the past few years, such targets are not consistent with many important theoretical considerations and experimental data, and that these considerations and data are most easily explained by a model of speech production whose only invariant targets are auditory perceptual targets. The remainder of this article makes this case in four parts. First, we posit that direct, accurate feedback concerning the locations and degrees of key constrictions in the vocal tract is not generally available to the central nervous system in a form suitable for movement planning. Such information appears to be crucial to the formation of a constriction representation for speech movement planning, so its absence poses a great difficulty to constriction target theories. In contrast, auditory perceptual feedback is readily available to the central nervous system. Second, the observation of approximate invariance in constriction location and degree seen during normal vowel production is addressed. This observation might appear to be evidence for invariant constriction targets. However, we show how approximate invariance in constriction location and degree can arise in control systems that do not use invariant constriction targets. Furthermore, such a system maintains a higher degree of motor equivalence than systems utilizing invariant constriction targets. This leads to the third part of our treatment, where we claim that invariant constriction targets would unnecessarily limit the motor equivalent capabilities of the speech motor system and are incompatible with recent experimental data from Savariaux et al. (1995a) and Perkell et al. (1993, 1994). Finally, we claim that American English /r/, which is often produced with two completely different constriction patterns by the same speaker in different contexts, is strong evidence against invariant constriction targets, instead indicating that the only invariant targets are of an acoustic or auditory perceptual nature. In this article we limit our claims to vowels and semivowels, although we suspect that these same claims may hold true for all speech sounds.

2. Unlike auditory perceptual feedback, direct, accurate feedback about constrictions is not generally available to the central nervous system

The first part of the argument against invariant constriction targets for vowels and semivowels concerns the observation that information about the shape of the vocal tract is not directly available to the central nervous system in the form of vocal tract constriction locations and degrees. This is not to say that constriction information cannot in principle be *derived* from available sensory information with

appropriate processing. Instead, we argue that available sensory information *in its raw form* is not organized in a constriction reference frame suitable for the planning of speech movements, thus necessitating a learned neural mapping between the sensory representations and a neural representation in a constriction reference frame to be used for movement planning. We will further argue that learning such a mapping is made very difficult, if not impossible, by the lack of an appropriate "teaching signal" and the many-to-one and one-to-many relationships between available sensory information and constriction parameters. Finally, we suggest that the closest thing to a teaching signal for learning constriction parameters is probably the acoustic signal after transduction by the auditory system, which is actually feedback in an auditory perceptual reference frame. If this feedback were used as a teaching signal to learn the required mapping from sensory information into a "constriction" planning frame, the resulting planning frame would be better characterized as an auditory perceptual planning frame rather than a constriction planning frame. Mappings of this type are easily learned by neural networks, as evidenced by the success of recent neural network models utilizing acoustic/auditory spaces for movement planning for vowels (Bailly et al., 1991, 1997; Guenther, 1995b).

Any neural controller that hopes to reach constriction targets must have accurate information about the position of the vocal tract within this constriction coordinate frame. For example, controllers that use some sort of feedback representation of the current vocal tract shape in order to produce movements that zero the difference between the current position and the target position in constriction space (e.g., Guenther, 1994, 1995a; Saltzman and Munhall, 1989) rely on the accuracy of this feedback information. Even a purely feedforward controller must somehow know which muscle length commands to issue in order to achieve a desired constriction target, thus requiring knowledge of the relationship between muscle lengths and vocal tract shapes within the constriction frame.

The main sources of information concerning vocal tract shape are the outflow commands to the speech articulators, tactile and proprioceptive feedback from the speech articulators, and the auditory representation of the acoustic signal produced by the vocal tract. Many motor control models posit a role for an efference copy of the outflow command to the muscles (also referred to as "corollary discharge"). If one assumes that the outflow command to the muscles controlling the positions of the speech articulators provides an accurate representation of the position of the vocal tract in constriction space, however, one begs the question of how the controller knew the appropriate mapping from constriction targets to muscle commands in the first place. The relationship between muscle lengths and constriction locations and degrees is a complex one that differs significantly from individual to individual because it depends heavily on the sizes and shapes of the speech articulators and the locations of muscles within the articulators. This issue will be addressed further shortly, but for now it suffices to note that the relationship between the constriction and muscle length reference frames cannot be genetically encoded and must instead be learned by the nervous system. It follows that outflow commands to the speech articulators cannot provide accurate constriction informa-

tion unless they are first "tuned", presumably using some other accurate representation of constriction location and degree originating from either orosensory feedback or the auditory representation of the acoustic signal produced by the vocal tract.

A vast amount of tactile and proprioceptive feedback from the speech articulators is available to the central nervous system. Furthermore, the ability to compensate for constraints on the articulators such as bite blocks, even before the first glottal pulse (Lindblom, Lubker, and Gay, 1979), strongly implicates this orosensory information in the control of speech movements. The relevant question here, however, is whether orosensory feedback in its raw form is sufficient to provide accurate constriction location and degree information to the regions of the brain controlling speech production.

To a first approximation, proprioceptive feedback from muscle spindles provides information about muscle lengths and shortening velocities modulated by gamma motoneuron activity (e.g., Brooks, 1986; Gordon and Ghez, 1991; Matthews, 1972). The natural reference frame for this information is, therefore, a muscle length reference frame. The presence of muscle spindles in the tongue and other speech articulators has been known for some time (e.g., Cooper, 1953). As mentioned above, the problem for constriction theories concerning feedback in a muscle length reference frame is that the relationship between this information and the locations and degrees of vocal tract constrictions is complex. To see this, consider the task of determining the degree of the tongue tip constriction given the lengths of the tongue and jaw muscles. For the sake of illustration, assume that a single muscle controls the height of the jaw, a second muscle controls the height of the tongue body with respect to the jaw, a third muscle controls the front/back position of the tongue body on the jaw, and two more muscles determine the height and front/back position of the tongue tip with respect to the tongue body. (This is of course a gross oversimplification of the actual situation but suffices for the current point.) Given only the lengths of these muscles, it is impossible to determine the tongue tip constriction location and degree since this depends on the exact shape of the jaw, tongue, and hard palate of the individual. Furthermore, the relationship between the muscle lengths and the tongue tip constriction location and degree is many-to-one; e.g., different lengths of the jaw height muscle can be compensated by changes in the tongue body height muscle and/or the tongue tip height muscle to maintain the same constriction location and degree. Additional complications arise because equal-sized changes in any given muscle's length cause different-sized changes in constriction parameters depending on where in the vocal tract the tongue lies; in other words, the relationship between muscle lengths and constriction parameters is nonlinear. In summary, then, the relationship between muscle spindle feedback and the constriction reference frame is many-to-one, nonlinear, and dependent on the specific shape of the articulators in an individual. It is therefore clear that without further processing, muscle spindle feedback is not organized in a constriction reference frame, and the properties of any neural subsystem that might perform this further processing must be learned rather than genetically encoded since they must differ across individuals and must change as

an individual grows. Again, this suggests the need for a "teaching signal" that provides accurate feedback in the constriction reference frame.

Tactile feedback from mechanoreceptors in the speech articulators provides information about the locations of contact between the surfaces of articulators. It is likely that tactile feedback provides important information to the central nervous system for stop consonants and fricatives, where complete or near-complete closure of the vocal tract is required. However, serious problems arise when one considers the task of deriving constriction parameters from tactile information for vowels and semivowels. First, for low vowels in particular, the relationship between a given pattern of tactile stimulation and the resulting degree of constriction can be one-to-many. Figure 2 shows a sketch from Stevens and Perkell (1977) of a coronal section through the vocal tract for different vowels. Consider the mid-vowel and low-vowel cases in Figure 2. Depending on the shape of the tongue (e.g., whether the midsagittal portion "peaks up" or "peaks down"), the same pattern of contact can correspond to different constriction degrees, thus making it impossible even in principle to determine constriction degree accurately given only the pattern of contact. This is particularly problematic in the low-vowel case, where there is little or no contact between the tongue and the palate for a wide range of constriction sizes. Second, the relationship between the pattern of contact and the constriction degree can also be many-to-one. That is, depending on the shape of the tongue, several different tactile patterns can all correspond to the same constriction degree. This observation holds not only for vowels and semivowels, but for fricatives and stop consonants as well. Given these considerations, we conclude that tactile information in its raw form does not accurately and uniquely specify constriction parameters, and again further learned processing would be needed to derive this information (to the degree that this is even possible) from tactile patterns.

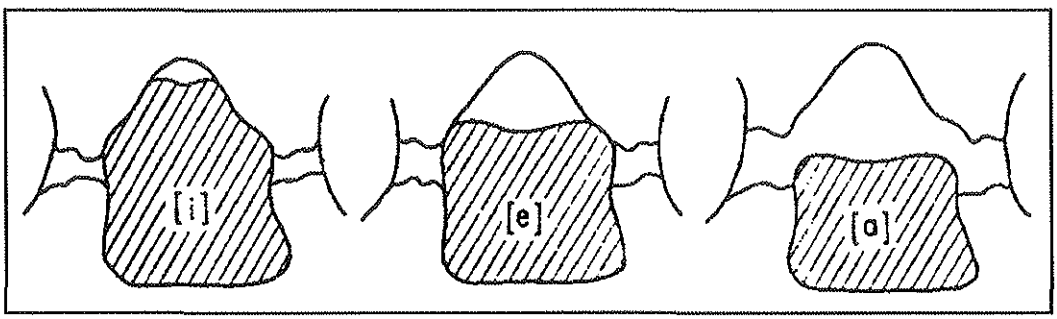

FIGURE 2. Sketches of a coronal section through the vocal tract for a high vowel (left), mid vowel (center), and low vowel (right). [Reprinted from Stevens and Perkell (1977).] The hatched areas represent the tongue. The size of the vocal tract constriction depends not only on the pattern of contact of the tongue with the teeth and hard palate but also on the shape of the tongue dorsum in the coronal plane, particularly in the low and mid vowel cases. This illustrates that tactile feedback alone does not uniquely specify the size of the vocal tract constriction.

Still, it seems highly likely that some combination of efference copy, tactile, and proprioceptive information plays an important role in relaying the state of the vocal tract to the central nervous system for the control of ongoing speech movements. In a review of research on various forms of feedback interruption in speech, Borden (1979) concluded that internal feedback of outflow commands likely plays an important role in normal speech. MacNeilage, Rootes, and Chase (1967) describe a patient with severe orosensory deficits but no known motor or speech perception deficits who's speech was essentially unitelligible. Although this suggests that orosensory information plays an important role in the development of speaking skills, it does not rule out the possibility that orosensory information is used only for development and not for the control of ongoing speech. Other researchers have investigated this possibility by temporarily inhibiting orosensory feedback in normal speakers. Lindblom, Lubker, and McAllister (1977) reported that temporary disruption of tactile information from labial and oral mucosa greatly impaired compensatory articulation in bite block speech (see also Hoole, 1987). Borden, Harris, and Oliver (1973) showed that mandibular sensory nerve blocks significantly decreased the intelligibly of speech produced by some (but, interestingly, not all) of their subjects. Analogous results have surfaced in the arm movement control literature. In studies of patients who developed severe proprioceptive deficits in their upper extremities after childhood, Ghez, Gordon, and Ghilardi (1995) and Gordon, Ghilardi and Ghez (1995) noted that although these deafferented subjects could roughly reach toward targets, their movements were very inaccurate when compared to normal subjects. The deafferented subjects' errors were consistent with the hypothesis that proprioceptive information is needed to allow compensation for the inertial properties of the limb. These results led the researchers to conclude that proprioceptive information is used to update an internal model of the limb's properties that is necessary for accurate reaching.

Probably the most accepted view in the motor control literature of the role played by outflow command, tactile, and proprioceptive information in movement control concerns the notion of an "internal model"; e.g., a learned neural mapping from information in a frame closely related to the positions of articulators into the reference frame for movement planning (e.g., a constriction frame or auditory perceptual frame for speech movements). Such a mapping has been termed a "forward model" by Jordan (1990) and has been used in different capacities in adaptive models of speech production (e.g., Bailly et al., 1991; Guenther, 1994, 1995a,b) and other motor tasks such as reaching (e.g., Bullock, Grossberg, and Guenther, 1993; Jordan, 1990). A typical neural network construct for learning a forward model is illustrated in Figure 3.

Current models which include a forward modeling component rely on a teaching signal to guide learning. In essence, the teaching signal provides the forward model with the output it should produce given the current inputs. Later, the forward model's output can be used in place of the teaching signal to identify the current location of the vocal tract in the planning reference frame.

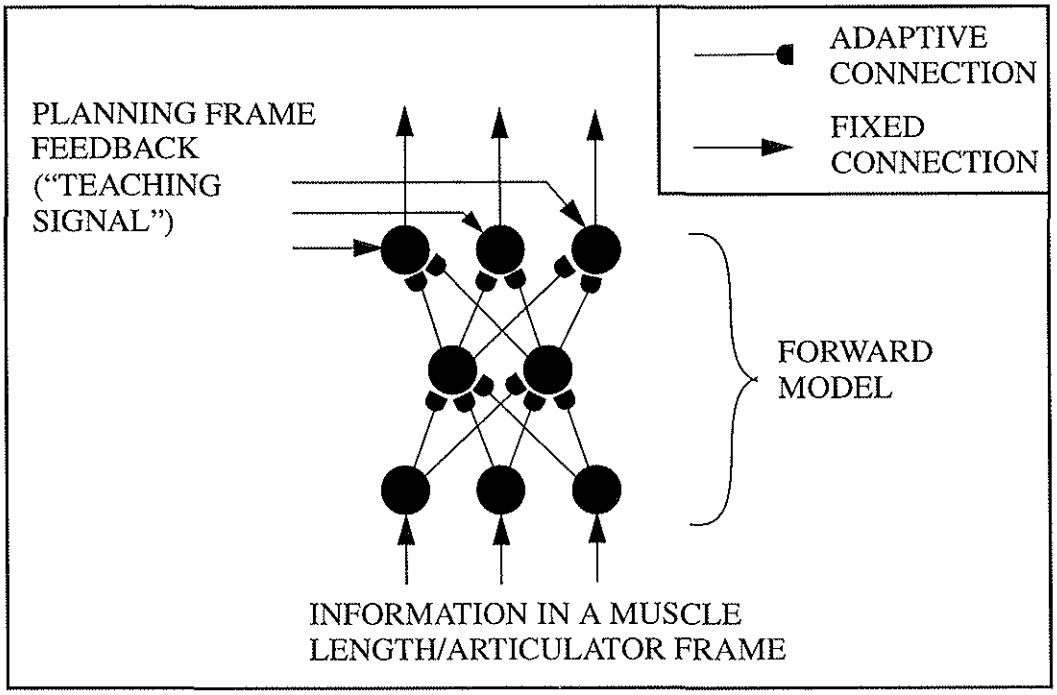

FIGURE 3. Typical neural network construct for learning a forward model. Black circles represent network "cells" or nodes, arrows represent synaptic connections whose strengths do not change, and filled semicircles represent adaptive synapses. The output stage of the forward model receives environmental feedback representing information in the planning reference frame. The input stage receives information about the current position of the articulators in a muscle length or articulator reference frame. This information can come from orosensory feedback or an efference copy of the outflow commands to the muscles. The planning frame feedback acts as a "teaching signal" that allows the forward model to learn the mapping between the articulator and planning reference frames by changing the strengths of the adaptive synapses. The learned forward model can then be used in place of planning space feedback from the environment for ongoing movement control (e.g., Bullock, Grossberg, and Guenther, 1993; Guenther, 1994, 1995a,b) or to train an inverse model to control the articulators (e.g., Bailly et al., 1991; Jordan, 1990).

An example of the forward modeling approach occurs in the DIVA model, schematized in Figure 1. In this case, a forward model that transforms information about the positions of articulators into an auditory perceptual reference frame is learned as follows. During babbling, articulator positions commanded by the system lead to an acoustic signal[2]. The articulator positions (available through outflow commands or orosensory feedback) act as the input to the forward model (see Figure 3). At the same time, the auditory system transduces the acoustic signal, resulting in an auditory perceptual representation that acts as the teaching signal for the forward model. The adaptive weights in the neural network are adjusted so that the forward model learns to match its output to the teaching signal given its current articulator position input. After learning, the forward model can be used in place of auditory feedback to indicate the current state of the vocal tract in the planning reference frame (i.e., the auditory perceptual frame) in order to determine

which commands to issue to the articulators to reach the current auditory perceptual target. That is, the model can work in the absence of auditory feedback once the forward model has been learned.

This example indicates how the nervous system could learn a forward model that encodes the relationship between orosensory feedback and the corresponding auditory signal by using a teaching signal available through auditory feedback during babbling. The forward model construct, however, appears to be insufficient for explaining how accurate information concerning constriction locations and degrees could be obtained by the central nervous system. The problem is that, unlike the auditory perceptual forward model, no teaching signal is available to accurately signal the locations and degrees of key vocal tract constrictions so that the neural mapping from orosensory feedback to constriction parameters can be learned. Perhaps the closest thing to this kind of teaching signal is the acoustic signal produced by the vocal tract after transduction by the auditory system. This is because of the relatively strong correlation between acoustic information and constriction locations and degrees (e.g., Coker, 1976). However, a forward model trained using this teaching signal is clearly better characterized as an auditory perceptual forward model rather than a constriction forward model.

Furthermore, it is unlikely that a forward model whose output is in constriction coordinates could self-organize in the absence of a teaching signal. Current self-organizing neural network architectures generally work by extracting statistical regularities in their training inputs (e.g., Grajski and Merzenich, 1990; Grossberg, 1976, 1980; Guenther and Gjaja, 1996; Kohonen, 1982; Sutton, Reggia, Armentrout, and D'Autrechy, 1994; von der Malsburg, 1973). As described above, however, constriction size is a very complex function of information in tactile and articulator reference frames. The many-to-one and one-to-many aspects of this function imply that it is not represented simply by regularities in the statistical distribution of the input information, and thus the constriction representation could not be extracted by these networks. Of course, the lack of an existing neural network architecture that can extract accurate constriction information without a teaching signal does not imply that it is *impossible* to self-organize such a representation, but our current understanding of how neural mappings are learned, coupled with the complexity of the mapping in question, certainly speaks against the plausibility of such a self-organizing process.

2. Of course, auditory feedback is not available for all configurations of the vocal tract (e.g., during stop closure). A key property of learned mappings implemented by neural networks is that they can generalize their performance to inputs which they did not encounter during learning. The forward model in Figure 1 is trained by matching articulator configurations to the auditory perceptual space values produced by these configurations during babbling. When the vocal tract configuration contains a stop closure during babbling, no auditory information is available and no learning occurs. Whenever the model produces the same configuration during performance, however, the forward model generates auditory perceptual space values due to the generalization property. This generalization effectively provides the smooth extrapolation of the internal representation of auditory space into regions where auditory information is unavailable.

To summarize, this section has outlined why it would be difficult, if not impossible, for the nervous system to derive an accurate representation of constriction location and degree from available sensory information. In contrast, it is relatively easy to see how an auditory perceptual representation can be derived, either directly from auditory feedback (during development) or indirectly from outflow command, tactile, and proprioceptive information processed by a forward model trained using auditory feedback during development.

3. Approximate invariance of constriction locations and degrees can arise in controllers which do not utilize constriction targets

Just as a constriction target may correspond to a range of articulator configurations, an auditory target may correspond to different sets of vocal tract constrictions. This implies that a speech production system whose only invariant targets are auditory targets will be capable of greater flexibility in selecting the final vocal tract configuration than a speech system based on invariant constriction targets. It does not, however, imply that a system with auditory targets must exhibit larger variability during unconstrained speech. To illustrate this fact, we present in this section a controller that uses invariant auditory targets but consistently tends toward a preferred vocal tract configuration when many possible configurations produce the desired acoustic output.

The relationship between auditory perceptual variables and articulator positions can be characterized as follows:

$$x = f(\theta) \tag{4}$$

where x is a vector specifying the position in auditory perceptual space, θ is a vector specifying the position in articulator space, and the function $f(\ )$ is the nonlinear mapping between these spaces. In the current case, x is a three-dimensional vector whose components are the Miller auditory perceptual dimensions defined in Equations 1 through 3, and θ is a seven-dimensional vector defining the positions of the seven articulators in the Maeda articulatory model.

In order to follow auditory perceptual trajectories in a manner that does not associate an invariant vocal tract shape target to every invariant auditory perceptual target, the DIVA model maps from desired movement *directions* (or, more precisely, velocities) in auditory perceptual space into articulator velocities that carry out these desired auditory perceptual trajectories. Such a mapping can be characterized mathematically by first taking the derivatives of both sides of Equation 4:

$$\dot{x} = J(\theta)\dot{\theta} \tag{5}$$

where J(θ) is the Jacobian of the function $f(\theta)$, then inverting this equation:

$$\dot{\theta} = G(\theta)\dot{x} \tag{6}$$

where G(θ) is a generalized inverse, or pseudoinverse, of the Jacobian matrix. Given that there are redundant degrees of freedom in the articulator set, J is not invertible and G must be one of the many possible generalized inverses of J. The choice of generalized inverse can affect the behavior of the system.

The most common choice of pseudoinverse, the Moore-Penrose (MP) pseudoinverse, results in a controller that selects the smallest movement in articulator space that will produce the desired movement in planning space. However, difficulties arise from the selection of this inverse, as Klein and Huang (1983) and Mussa-Ivaldi and Hogan (1991) discuss in the context of the control of a robotic arm. In particular, the MP pseudoinverse does not produce the same articulator configuration each time it returns to a given point in the planning space. If a closed loop in planning space is traversed repeatedly by an arm controlled using the MP pseudoinverse, the result can be a consistent shift in joint angles which can drive the system to the extreme limits of its joint ranges and leave the arm "curled up" in an unnatural position. Similarly, for a speech system based on auditory targets, this property of the MP pseudoinverse can result in different constrictions across utterances of the same phoneme and, after several repetitions of the same sound pattern, can curl the articulators into an awkward or extreme articulator configuration.

In contrast, neither the reaching motor system nor the speech motor system is characterized by such behavior. Psychophysical studies of reaching and pointing tasks imply a degree of invariance in the motor system, such that repeated execution of a pointing or reaching task generally results in a similar final posture of the arm across trials. Studies of pointing movements with the elbow fully extended indicate that the final posture of the arm is relatively invariant for a given target position (Hore, Watts, and Vilis, 1992; Miller, Theeuwen, and Gielen, 1992). For pointing movements on a planar surface, Cruse, Brüwer, and Dean (1993) report that the final postures "were virtually independent of the configuration at the start of the pointing movement" (p. 131), and for reaches to grasp an oriented object, Desmurget et al. (1995) similarly report that "the final limb angles were highly predictable" (p. 905). Although the final postures of unconstrained, three-dimensional reaches to a given target did show a dependence on starting configuration in the Soechting Buneo, Herrmann, and Flanders (1995) paradigm, the extent of this variability in comparison to the total variability possible given the geometry of the arm was not addressed. Relative invariance was addressed in the more restricted paradigm of Cruse (1986), where it was found that the range of configurations reached was very limited in comparison with the range physically possible for completing that task. It therefore appears that although some variability in final posture is seen, the motor system uses a far smaller range of final postures than is possible given the redundancy of the arm.

The existence of approximate constriction invariance[3] in the production of phonemes has not been so thoroughly investigated, but it is a common assumption historically that vocal tract constrictions are approximately invariant for a given phoneme across utterances. In fact, vowels are typically defined by the locations and degrees of their constrictions. For example, /i/ is identified as a high front vowel in reference to a tight constriction (high tongue position) formed at the front of the vocal tract. Of course, the close relationship between vocal tract constrictions and the resulting acoustic output implies that constriction locations and degrees must be somewhat consistent across utterances. The relevant question here, however, is whether this consistency is greater than is necessary to produce recognizable acoustic output. Figure 4 addresses this question. Figure 4a shows the range of articulator configurations of the Maeda articulator model that can be used to produce acceptable tokens of the vowel /ɛ/. This figure was created by superimposing the Maeda articulator configurations that produced formant frequencies F1 and F2 within +/-25 Hz and F3 within +/-50 Hz of "ideal" values for the vowel /ɛ/ (F1 = 530 Hz, F2 = 1840 Hz, F3 = 2480 Hz). Vocal tract outlines obtained from x-ray tracings of a speaker pronouncing the vowel /ɛ/ in four different contexts (/hənɛ/, /həkɛ/, /həpɛ/, /hədɛ/) are overlayed in Figure 4b. The variability evident in the speaker's utterances is quite restricted compared to the possible range of configurations shown in Figure 4a, with the tongue shapes confined to somewhere near the midrange of possible configurations for the vowel.

One approach to reproducing this approximate constrictional invariance in a pseudoinverse-style controller is to use an integrable pseudoinverse like the one presented by Mussa-Ivaldi and Hogan (1991). Unless a perturbation or constraint is introduced during speech, this pseudoinverse maps each position in planning space to a unique articulator configuration, although the system may still adopt unusual configurations to compensate for constraints. However, once an unnatural configuration is established, integrability of the pseudoinverse will ensure that it is preserved. This is a problem because even awkward or extreme configurations are maintained.

Another approach that avoids the problem of maintaining awkward postures is to bias the system so that it always chooses movements that lead toward more comfortable configurations. That is, from the infinite number of possible articulator velocity vectors $\dot{\theta}$ that move the vocal tract in the desired auditory space direction $\dot{x}$, we can choose one that also moves the articulators toward the centers of their ranges as much as possible[4]. This property, which we will refer to as *postural*

3. We are not claiming that constrictions are invariantly produced during speech. By "approximate constriction invariance" we simply mean that under normal conditions the speech production system uses a limited set of the articulator configurations that could in principle be used to produce a given speech sound. The variable aspects of speech production are numerous and have been the subject of a large number of studies (e.g., see Perkell and Klatt, 1986). Guenther (1995a) describes a convex region theory of the targets of speech that is implemented in the current model and provides an account for many aspects of articulatory variability, including motor equivalence, contextual variability, carryover coarticulation, anticipatory coarticulation, and variability due to changes in speaking rate.

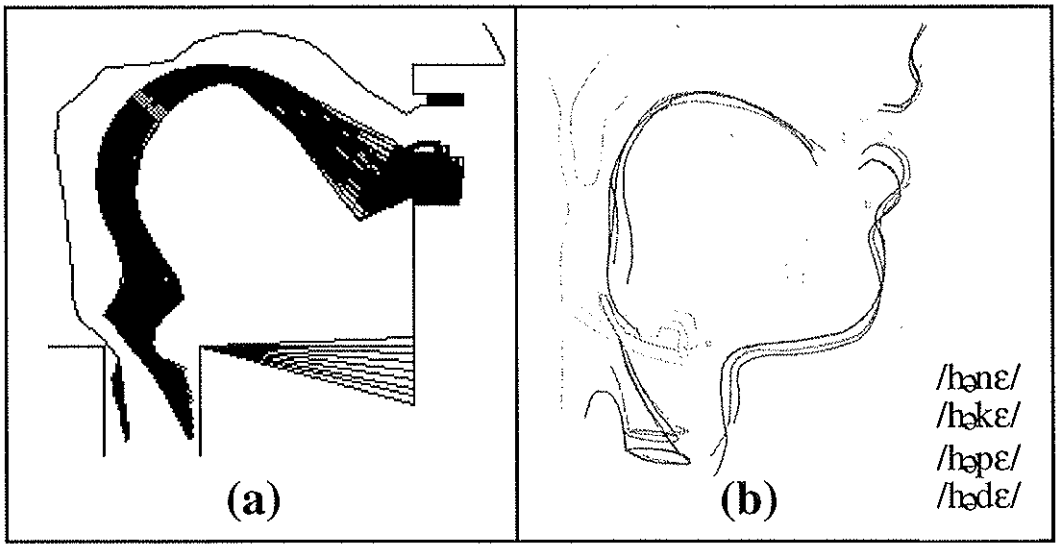

FIGURE 4. (a) Possible configurations for producing /ɛ/ using the Maeda articulatory synthesizer. (b) Configurations used by a speaker to produce /ɛ/ in four different consonant contexts. [Data courtesy of Joseph Perkell.]

relaxation, can be implemented with a modified differential controller based on the equation:

$$\dot{\theta} = G(\theta)\dot{x} + R(\theta) \tag{7}$$

The vector $R(\theta)$ will be referred to as a *stationary* vector because it specifies a movement of the articulators that does not affect the position in auditory planning space. Such vectors exist due to the redundancy of the mapping from the planning space to the articulator space. Given an explicit expression for the Jacobian, a stationary vector $R(\theta)$ which biases the system toward a comfortable posture can be calculated by taking a vector in the direction of the most comfortable posture and projecting it onto the nullspace of J. This approach has been described in the robotics literature (e.g, Liégeois, 1977; Baillieul, Hollerbach, and Brockett, 1984).

Unfortunately, the robotics approach embodied by Equation 7 cannot generally be applied to speech systems, either biological or artificial, that use auditory planning spaces. This is because the mapping between articulator configurations and acoustic output, $f(\theta)$ in Equation 4, is not generally known in a form that allows explicit computation of the Jacobian matrix or a generalized inverse of the Jacobian. In an artificial system, the acoustic output is usually calculated by applying digital signal processing techniques to determine the sound waveform that would be produced by a series of tubes approximating the shape of the vocal tract when excited by an

4. It is quite possible that the elastic and compressive properties of the tissues making up the speech articulators would provide a natural tendency for the articulators to move toward more central configurations, similar to the biasing force described here and learned by the model.

acoustic energy source such as vocal fold vibration or fricative noise (e.g., Rubin et al., 1981). This approach does not provide one with a simple formula for $f(\theta)$ that can be used to calculate the Jacobian matrix. The situation is even worse in biological speech motor control systems since the Jacobian matrix depends on the sizes and shapes of the speech articulators and thus differs from individual to individual and changes as an individual grows.

An alternative approach is to use an adaptive controller which *learns* appropriate values of $G(\theta)$ and $R(\theta)$ in Equation 7 during babbling, thus eliminating the need for an explicit formulation of the Jacobian. This is the approach used by the DIVA model. The basic idea of this learning process is sketched out here and detailed for an arm movement control model in Guenther and Micci Barreca (1997).

A directional mapping between auditory perceptual space and articulator space approximating Equation 7 is learned by the DIVA model during a babbling cycle in which random movements of the speech articulators, $\Delta\theta_B$, result in changes in the acoustic signal produced by the vocal tract. This changing acoustic signal is perceived by the model as a trajectory in auditory perceptual space, and the direction of the trajectory in auditory space is mapped through the (initially inaccurate) directional mapping to form a predicted value of the babbling movement, $\Delta\theta$, that caused the sound. The difference between the babbled movement and the predicted movement acts as an error signal that is used to update the directional mapping.

The learned directional mapping consists of approximations to the generalized inverse G and the stationary vector R in Equation 7. These entities are learned using two sets of radial basis functions (RBFs). One set learns the entries of G by minimizing the cost function:

$$H_1 = \sum_i (\Delta\theta_{Bi} - \Delta\theta_i)^2 . \tag{8}$$

This cost function is simply the square of the error described above, and it causes the system to learn an approximation to a generalized inverse G that minimizes the error between the babbled movement direction and the movement direction predicted by the directional map.

The second set of RBFs learns the components of $R(\theta)$ by minimizing the cost function:

$$H_2 = \sum_i (\Delta\theta_{Bi} - \Delta\theta_i)^2 + \beta_i \left(\frac{\theta_i - \theta_i^c}{\theta_i^r} \right)^2 \tag{9}$$

where θ_i^c and θ_i^r are the central position and range of motion of the i^{th} articulator. The first part of this cost function has the effect of keeping the learned vector R from producing movements of the articulators that would drive the system away

from the desired auditory trajectory. The second part has the effect of choosing values of R that move the articulators closer to the centers of their ranges. The RBF equations and derivations of the learning rules using these cost functions are provided in Appendix A.

Incorporation of this learned directional mapping in the DIVA model (see Figure 1) results in a system which, although flexible in overcoming articulatory constraints, otherwise tends consistently toward "comfortable" configurations. Figure 5 shows the results of simulations of the model with and without the postural relaxation component. For each vowel, the articulators were initialized to five different starting configurations, corresponding to the phonemes /p/, /k/, /d/, /n/, and /ə/. From each starting position, the model moves the articulators until the invariant auditory perceptual target for the vowel is reached. The resulting set of final configurations for each vowel were then superimposed to produce the panels of Figure 5. The left figure in each panel illustrates the variability in the absence of the postural relaxation component, and the right figure in each panel illustrates the reduced variability that occurs with the addition of the postural relaxation component. This figure clearly illustrates that the postural relaxation scheme has the effect of biasing the system toward a "canonical" vocal tract configuration for each phoneme even in the absence of an invariant vocal tract shape target.

It is very important to note that although the incorporation of postural relaxation reduces the variability exhibited during normal speech, it does not hinder the system's flexibility in overcoming constraints such as bite blocks or lip tubes. That is, the system will still automatically adopt unusual configurations when necessary to produce a given phoneme under constraining conditions. This important property, which distinguishes the current model from models that utilize invariant constriction targets or invariant articulator configuration targets, is discussed and illustrated in the next section. Thus, the model can account for both the approximate vocal tract shape invariance seen under normal speaking conditions and the ability to utilize new shapes (including changes in constriction location) when necessitated by externally imposed constraints on the speech articulators.

4. Invariant constriction targets would unnecessarily limit the motor equivalent capabilities of the speech production system

Perturbation studies (e.g., Abbs, 1986; Abbs and Gracco, 1984; Lindblom, Lubker and Gay, 1979; Savariaux et al., 1995a) have established that the speech production system exhibits motor equivalence by using new articulator configurations that preserve the perceptual identity of a phoneme when the default articulator configuration for that phoneme is not possible due to an externally imposed constraint such as a bite block or lip tube. The relevant question in the current context is whether speaker compensation is geared toward preserving a specific set of vocal tract constrictions or toward maintaining (in any manner possible, including

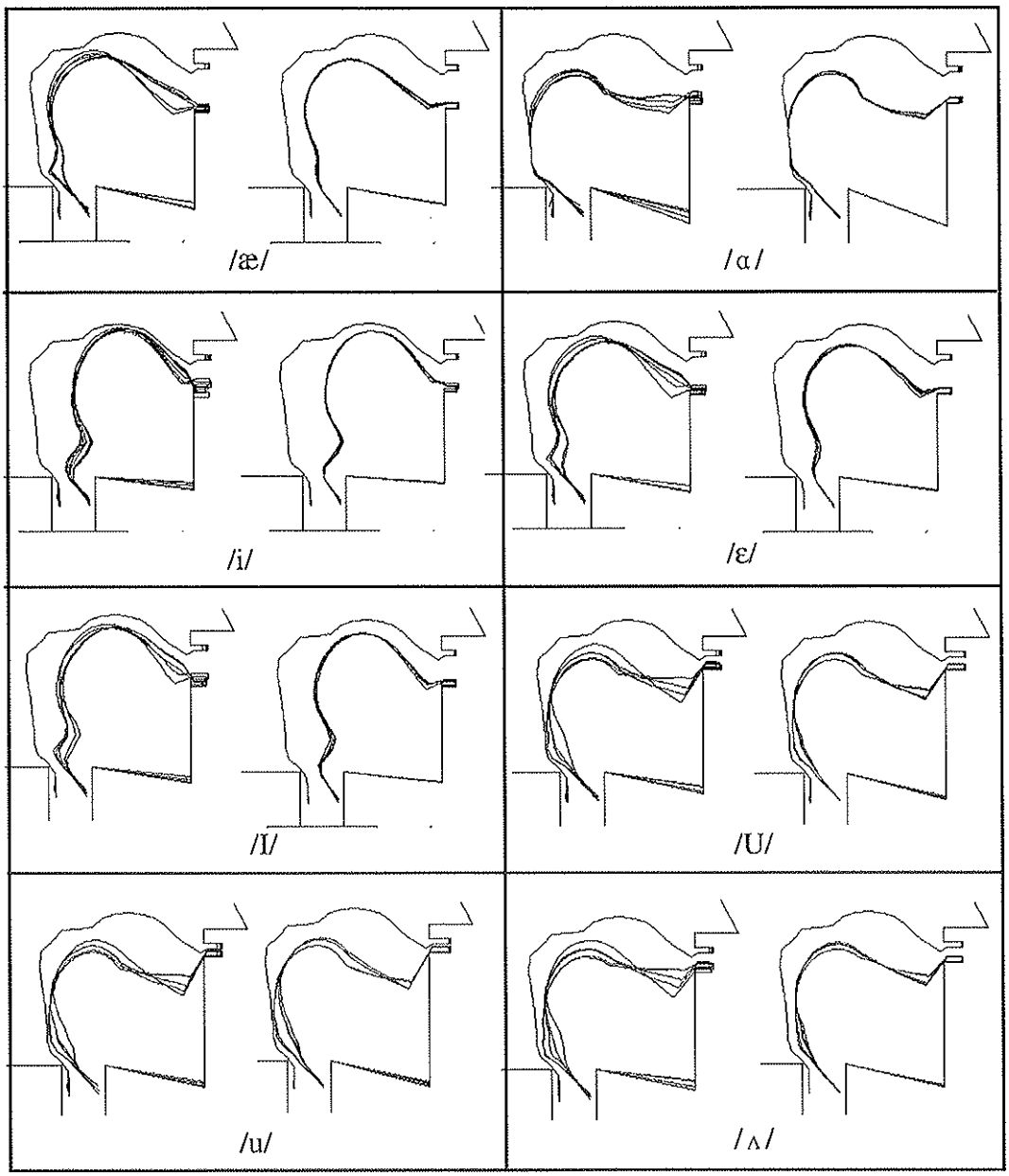

FIGURE 5. Reduction of articulatory variability with the addition of a postural relaxation component to the mapping between auditory perceptual space and articulator space. For each vowel, the left figure in the panel shows the variation in articulator configuration that arises in a pseudoinverse-style controller without postural relaxation when starting from five different initial configurations of the vocal tract (corresponding to the phonemes /p/, /k/, /d/, /n/, and /ə/). The right figure in each panel shows the reduced variability that occurs with the addition of the postural relaxation component.

changing the constrictions) the relevant auditory perceptual aspects of the sound being produced. The higher level of motor equivalence possible with the latter strategy would clearly be advantageous to the speech production system since the goal of speech is ultimately to produce recognizable phonemes. This section investigates data indicating that the speech motor system appears to take advantage of the additional motor equivalence that is possible when using invariant auditory perceptual targets rather than invariant constriction targets.

Before investigating recent motor equivalence studies, we will briefly illustrate the motor equivalent capabilities of the DIVA model when using acoustic space targets for vowels (Johnson and Guenther, 1995). Figures 6 and 7 show the results of simulations carried out with a version of the model that utilized point targets in F1/F2 space corresponding to typical formant values for each of 10 American English vowels (Rabiner and Schafer, 1978). Figure 6 illustrates the model's performance in the absence of constraints on the articulators. Typical values of F1 and F2 for the vowels are indicated by crosses, and values produced by the model when starting from a neutral vocal tract configuration are indicated by triangles. An ellipse is drawn around the typical value and the value produced by the model for each vowel. The model gets very close to the target for each vowel, although the /ɑ/ produced by the model is significantly further away from its target value than the other vowels. (This latter result appears to reflect a difficulty inherent to the Maeda vocal tract in reaching the typical /ɑ/ formants specified in Rabiner and Schafer, 1978.) Figure 7 illustrates the model's performance when the jaw is fixed in a position that is unnatural for most of the vowels, as would occur if a bite block were held between the teeth. Rectangles indicate formant values that would arise without compensation for the new jaw position, and lines connect these values to the typical values for the corresponding vowels. Despite the large formant shift induced by the bite block, the formant values produced by the model in the bite block condition are nearly identical to values produced in the unconstrained condition (Figure 6), indicating full compensation for the bite block. This compensation occurs automatically in the model, even though no training was performed with the jaw constraint present.

A striking example of the speech production system utilizing different constriction configurations to produce the same acoustic result is the American English phoneme /r/. Although clearly an instance of motor equivalence, this important case will be treated separately in the next section to allow a more detailed analysis.

Another form of motor equivalence involves trading relations between constrictions that achieve a relatively invariant acoustic result. Perkell, Matthies, Svirsky and Jordan (1993) studied correlations between the gestures of tongue body raising and lip rounding in subjects pronouncing the vowel /u/. These gestures were chosen because, although they affect the area function at different parts of the vocal tract, they play similar roles in shaping the acoustic spectrum, mainly decreasing the second formant. If a trade-off between the contributions from these two gestures is used to maintain the second formant in a constant range, the tongue height and lip rounding parameters should be negatively correlated. Three of the four

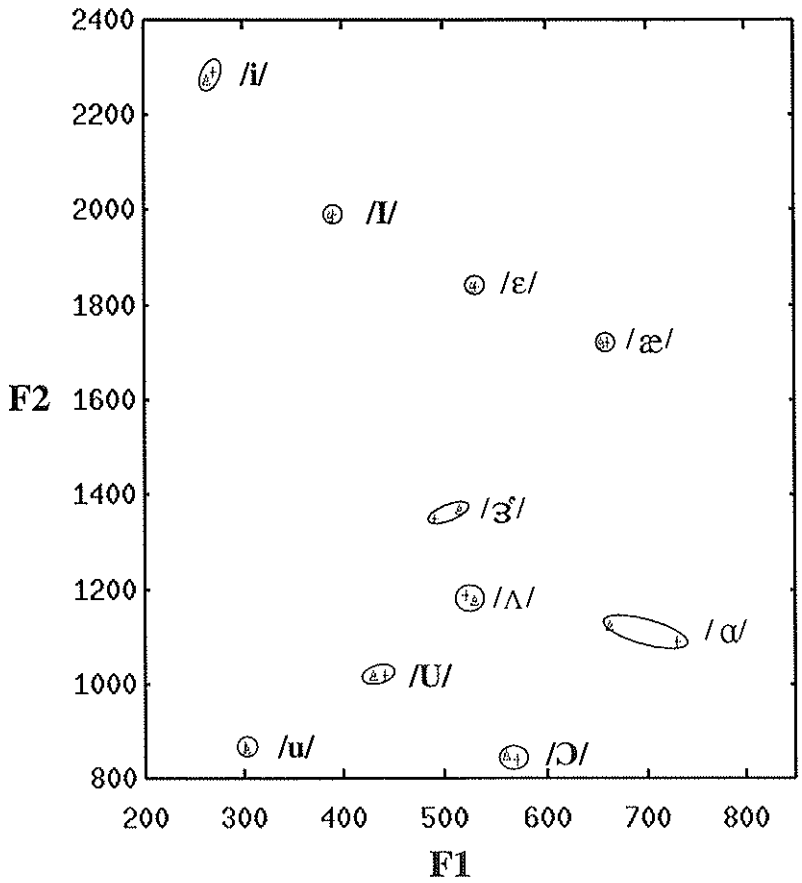

FIGURE 6. Typical values of F1 and F2 for American English vowels (indicated by crosses) and values produced by the model (indicated by triangles) when starting from a neutral vocal tract configuration. An ellipse is drawn around the typical value and the value produced by the model for each vowel.

subjects in the study did exhibit weak negative correlations, leading to the conclusion that their study provides "tentative support for motor equivalence at the area-function-to-acoustic-level" (p. 2960). A follow-up study of /u/, /r/, and /ʃ/ utterances (Perkell, Matthies, and Svirsky, 1994) found that the level of correlation between constriction parameters increased for less prototypical tokens. Since less prototypical tokens correspond to more vulnerable percepts, this supports the view that trading off between the two gestures is a mechanism for maintaining acoustic quality of the phoneme, and thus becomes more prominent as acoustic variability becomes less acceptable.

Because they show a variability in constrictions that acts to maintain an acoustic result, these data are very troubling to an invariant constriction theory of the targets of speech. However, they are easily captured by a theory in which the targets of speech production are multidimensional regions in auditory perceptual space, as in

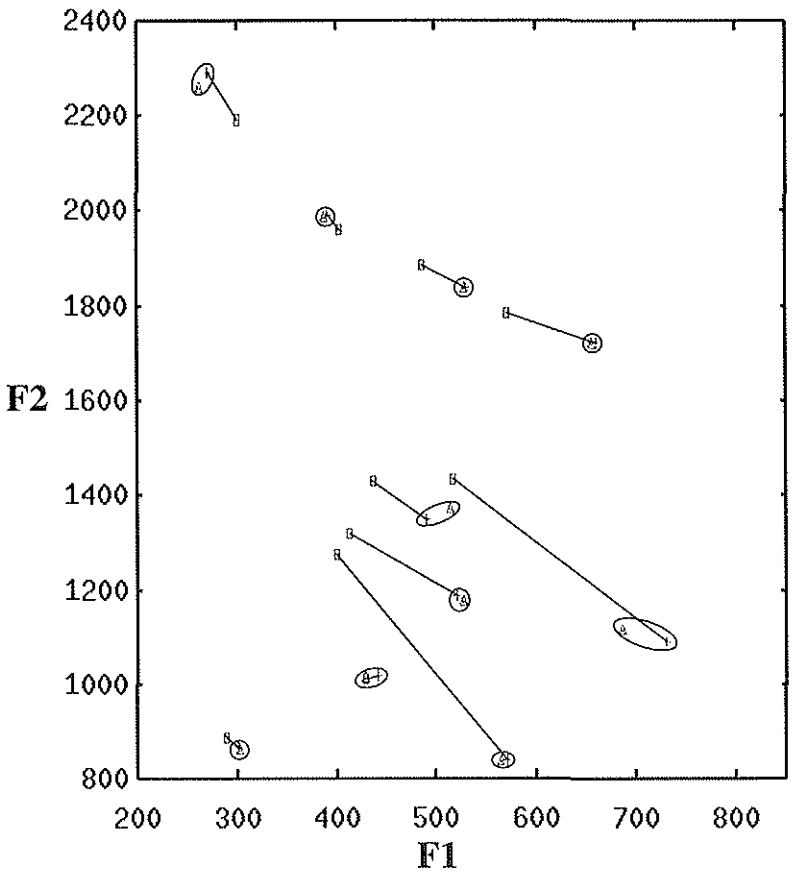

FIGURE 7. "Bite block" simulation in which the jaw parameter is clamped at a value of 0. Crosses indicate typical formant values for male speakers, triangles indicate the formant values produced by the model, and ellipses are drawn around corresponding typical values and model values. Rectangles indicate formant values that would arise without compensation. The formant values produced by the model in this condition are nearly identical to values produced in the unconstrained condition (Figure 6), indicating full compensation for the bite block. This compensation occurs automatically in the model, even though no training was performed with the jaw constraint present.

the current model. Guenther (1995a) illustrated how the model commands movements of the speech articulators only when needed to move the vocal tract to reach the target region of the current phoneme. If the system is already within the target region, no movements are commanded. A "prototypical" sound in the Perkell et al. (1994) study is presumably one where the production system has reached a position near the center of the target region and thus does not need to produce any compensatory movements. If, however, the tongue body constriction is lower than usual due to coarticulatory influences, the system will command movements including rounding of the lips to compensate until the edge of the auditory target region is reached, as seen in the Perkell et al. (1993, 1994) data.

Another study of motor equivalence involving constriction location was performed by Savariaux et al. (1995a,b; Savariaux, 1995). In this study, a lip tube perturbation was introduced to prevent speakers from forming the tight labial constriction associated with the French vowel [u]. Using nomograms obtained from Fant's model (Fant, 1992), Savariaux et al. (1995a) determined that the primary effect of the lip tube was a large increase in F2 and smaller increases in F1 and F3, and that compensation for the lip tube perturbation could be achieved by pulling the tongue body back to form a velo-pharyngeal constriction. Most of the speakers (7 out of 11) did "show an observable articulatory change, corresponding to a tongue backward movement, after the insertion of the tube between the lips" (p. 2440). These results clearly speak against an invariant constriction target for [u] in these speakers.

In contrast, compensatory changes in constriction location for the lip tube perturbation are predicted by models utilizing invariant auditory perceptual targets. This is illustrated by the results of a computer simulation of the lip tube study using the DIVA model (Figure 8). The auditory perceptual dimensions defined in the introduction were used to specify the target for [u]. Lip aperture was fixed at 4.9 cm^2 (the cross-sectional area of the lip tube used by Savariaux et al.), and the formant frequencies were adjusted to reflect the formant perturbations predicted by Savariaux et al. (1995a)[5]. The configurations used by one of the speakers from the Savariaux et al. (1995a) study who showed a change in articulator configuration after insertion of the lip tube are shown in the left half of Figure 8. The solid line indicates the configuration of the vocal tract without the lip tube. The dotted line in the left half of Figure 8 indicates that the subject showed some compensation in constriction location even on the first attempt. Similarly, the model immediately compensates for the lip tube by retracting the tongue body, as shown in the right half of Figure 8. To our knowledge, the current model is at present the only computational model of speech production that accounts for this immediate compensation in constriction location, i.e., without requiring additional learning after insertion of the lip tube perturbation.

Importantly, Savariaux et al. report that subjects generally did not move the tongue back as far as was predicted using the Fant (1992) model. This is evidenced by the fact that the resulting formant frequencies were generally higher than for normal productions. Still, perceptual tests of the lip tube utterances showed that 7 of the 11 lip tube [u]s were easily identified by listeners, even though the formant values for these [u] productions were typically somewhat higher than normal.

Why didn't subjects retract their tongues far enough to fully compensate for the formant shifts caused by the lip tube? Two explanations arise within the current

5. The formant perturbations produced by fixing the lip aperture alone differed significantly from the estimates provided by Savariaux et al. (1995a). We believe this reflects a limitation of the algorithm used to calculate formant values for extreme values of lip aperture in our model. To overcome this problem, the formant values calculated by the model under the lip tube condition were adjusted to reflect the same perturbations predicted by Savariaux et al. (1995) using Fant's (1992) model.

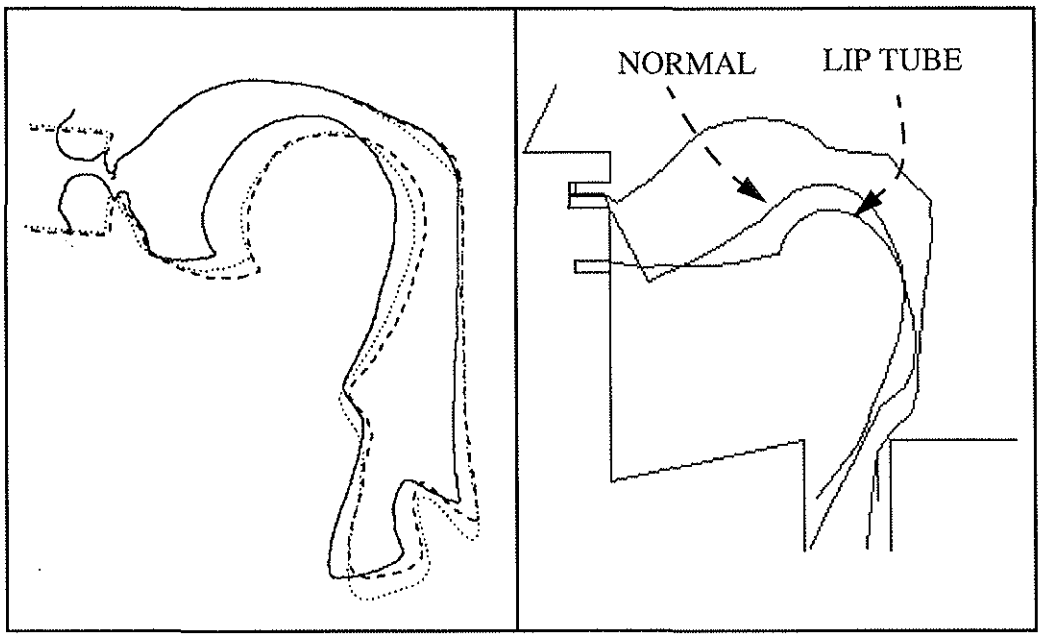

FIGURE 8. Vocal tract outlines indicating compensation for a lip tube perturbation by a subject from the Savariaux et al. (1995a) study (left) and by the DIVA model (right). The solid line in the left figure indicates the configuration of the vocal tract for [u] without the lip tube, the dotted line indicates the configuration on the first trial after insertion of the lip tube, and the dashed line indicates the configuration on the twentieth trial. This subject moved the tongue constriction further back in the vocal tract to compensate for the lip tube. The model's lip tube utterance shows a similar shift of the tongue body constriction, as shown in the right half of the figure. [Left half of figure adapted from Savariaux et al. (1995a).]

theory. First, the convex region theory implies that subjects will only compensate enough to get their productions into the acceptable region for [u], as described above for the Perkell et al. (1993, 1994) results. Thus it is no surprise that subjects' productions in the lip tube case do not show complete compensation in formant space. Second, formant frequencies are not the only perceptually relevant aspects of the acoustic signal. Subjects may have changed other aspects of the speech waveform to compensate in auditory perceptual space. The remainder of this section provides support for these explanations.

As described in the introduction, we have chosen the auditory perceptual dimensions suggested by Miller (1989) to represent auditory perceptual space in the model. Although the perceptual identification data of Savariaux et al. (1995b) are not easily separable in formant space, they are easily separable in the Miller auditory perceptual space. In fact, linear separability can be achieved using just two of the three dimensions, as illustrated in Figure 9.

There are at least two major advantages to Miller's auditory perceptual space over other formulations such as pure formants. First, ratios of formants provide some

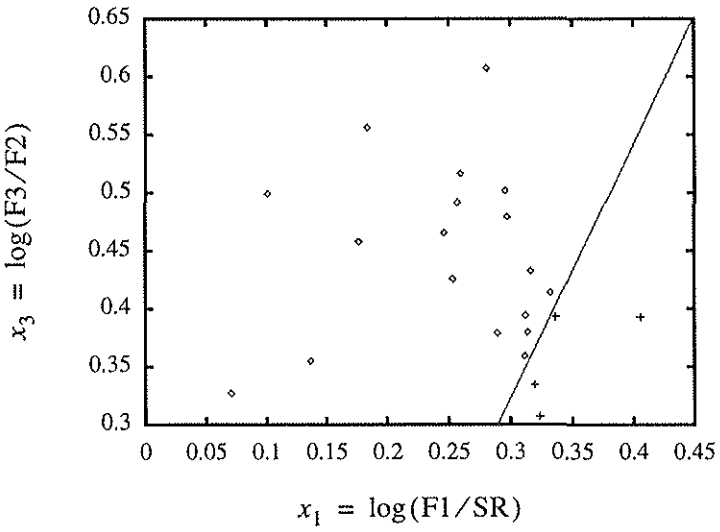

FIGURE 9. Perceptual recognition data of Savariaux et al (1995b) for normal and lip tube utterances, plotted in terms of two of the auditory perceptual dimensions of Miller (1989). Diamonds indicate easily identifiable [u] productions (>90% ID rate), and crosses indicate unidentifiable [u]s (<15% ID rate). A line has been drawn to illustrate the linear separability of these points in this auditory perceptual space.

normalization across speakers. For example, although the formants of women are generally higher than those of men for the same phoneme, the ratios are approximately the same across gender. The second advantage is that Miller's first dimension, x_1, incorporates information about pitch (recall that SR in Equation 1 is a function of F0), which appears to play an important role in speech perception.

As Savariaux et al. (1995b) point out, the view that speakers are modifying perceptually relevant aspects of the acoustic signal in addition to formant frequencies is supported by the tendency of speakers to raise their fundamental frequency when compensating for the lip tube perturbation. The mean increase in F0 across all subjects was 5.5 Hz (Savariaux, 1995). Figure 10 illustrates how each subject's F0 variation contributed to the location of his/her perturbed utterance in Miller's auditory perceptual space. The tails of the arrows in the figure indicate where a speaker's utterance would have fallen in the perturbed condition had he/she used the same F0 as in the normal condition. The arrow is then drawn to the actual perturbed utterance location. Note that several speakers raised their fundamental frequency just enough to shift their utterance from the region containing unidentifiable [u]s to the region containing easily recognizable [u]s.

It is interesting to note that this is another example of compensation taking place primarily for non-prototypical sounds (c.f Perkell et al., 1994), again providing supporting evidence that the targets of speech production are *regions*, not points. Figure 10 provides an illustration of the speech production system of several speakers compensating just enough to get into an acceptable region for the sound,

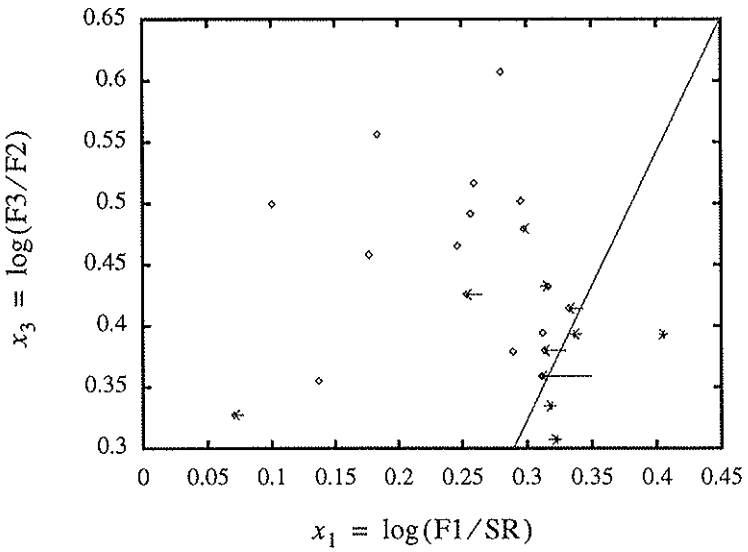

FIGURE 10. Investigating the compensatory effects of pitch changes in the perceptual recognition data of Savariaux et al. (1995) for normal and lip tube utterances. The line in the figure indicates the cutoff between acceptable [u] utterances (to the left of the line) and unacceptable [u]s (to the right). Arrows associated with lip tube utterances indicate how change in F0 contributed to the final utterance. The tails of the arrows indicate where a speaker's utterance would have fallen in the perturbed condition had he/she used the same F0 as in the normal condition. The arrow is then drawn to the actual utterance location. These data indicate that several subjects increased F0 just enough to "cross the line" into the acceptable region of auditory perceptual space for an [u], while those already in the acceptable region did not generally change F0. This is consistent with the convex region theory of the targets of speech (Guenther, 1995a) when extended to auditory perceptual space; see text for details.

as predicted by the convex region theory of Guenther (1995a) when extended to auditory perceptual space.

In summary, the results from the majority of subjects in the Savariaux et al. lip tube study are consistent with the idea that the invariant targets of speech production are multidimensional regions in auditory perceptual space. These results also contradict invariant constriction target theories, which cannot account for the compensatory change in constriction location seen in most subjects. Assuming that articulatory gestures such as tongue retraction are easier for some individuals than others, it is to be expected that the extent to which a single compensation strategy is expressed will vary from speaker to speaker. Therefore, it is not surprising that the degree of tongue motion and pitch increase varied across individuals in the Savariaux et al. study. Still, it is unclear why four subjects in the study did not manage to compensate perceptually for the lip tube perturbation. Several factors may have led to this inability. First, the lip tube may have simply made it physically impossible for these subjects to compensate given the specific shapes of their vocal tracts and articulators. Second, experience with the possibly awkward configurations required to produce compensation is likely very limited during normal

speech. This suggests that some subjects' "forward models" may have been inaccurate in these regions and might therefore have been incapable of supporting proper compensation. Both of these possibilities could explain why some speakers compensate while others do not. A third possibility is that these four subjects were using an invariant constriction or articulator target for [u] and therefore could not compensate for the tube. However, this possibility is inconsistent with the seven subjects who produced perceptually clear [u]s with the lip tube by pulling their tongue back and/or raising their fundamental frequency, thereby preserving perceptually relevant acoustic information. Although in conflict with the bulk of the data, further investigation within this paradigm is needed before ruling out the possibility that some speakers use invariant targets that are more "constriction-like" than the targets of other speakers.

5. The only invariant target for American English /r/ appears to be an acoustic or auditory target

Further evidence against constriction targets comes from studies of the American English phoneme /r/, which is a rare example of a phoneme for which very different articulations can produce very similar acoustic patterns. Furthermore, the same speaker will often use very different articulator configurations to produce /r/ in different contexts (Delattre and Freeman, 1968; Espy-Wilson and Boyce, 1994; Hagiwara, 1994, 1995; Narayanan, Alwan, and Haker, 1995; Ong and Stone, 1997; Westbury, Hashi, and Lindstrom, 1995). Figure 11 shows two such configurations for /r/, known generally as "bunched" and "retroflexed". Ong and Stone (1997) report a subject who used bunched /r/ in a front vowel context and retroflexed /r/ in a back vowel context. From this example, it is clear that phonetic context plays a major role in determining which variant of /r/ is used. Although the bunched and retroflexed variants are the most commonly reported, other investigators have suggested that more than two types are used, including Hagiwara (1994, 1995), who posits three variants, and Delattre and Freeman (1968) and Westbury et al. (1995), who suggest that a continuum of variants exist between extreme bunched and extreme retroflexed.

The existence of two or more completely different configurations for producing the same phoneme is difficult for theories that hypothesize invariant constriction targets. This is because the constriction locations and degrees used to produce the two /r/'s in Figure 11 are completely different (note particularly the tongue tip and tongue body constriction locations and degrees), so the corresponding targets must also be completely different. This leads to a rather unsatisfying explanation in which an individual chooses one or the other target depending on context. Although not completely unreasonable, this explanation is not very elegant. A more parsimonious explanation utilizing a single target specified within an acoustic or auditory perceptual planning frame is provided by Guenther (1995b) and is described in the following paragraphs. This explanation relies on two key characteristics of the current model: (1) the specification of phonemic targets and plan-

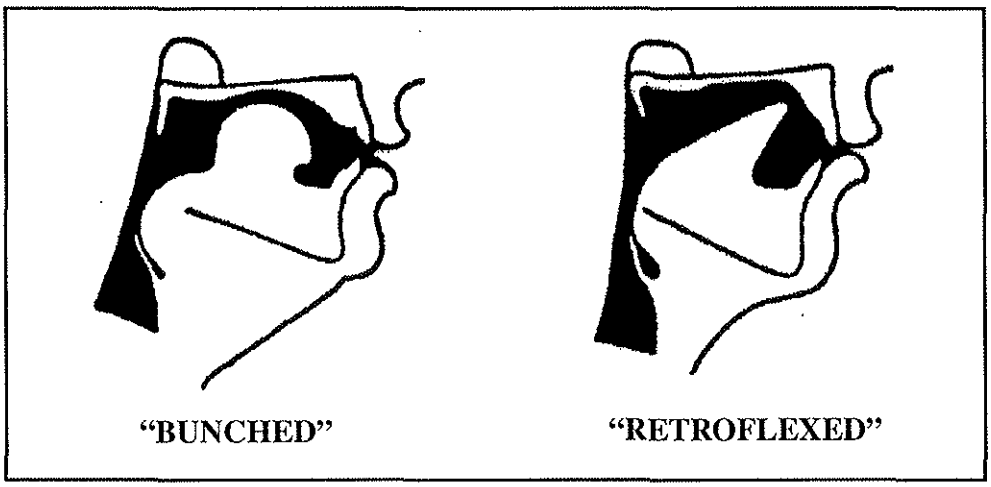

FIGURE 11. Two of the articulator configurations commonly seen during production of American English /r/ (after Delattre and Freeman, 1968). It has been noted by several investigators that the same speaker will often use two or more different vocal tract configurations to produce /r/ in different contexts (e.g., Delattre and Freeman, 1968; Espy-Wilson and Boyce, 1994; Hagiwara, 1994, 1995; Narayanan, Alwan, and Haker, 1995; Ong and Stone, 1997; Westbury, Hashi, and Lindstrom, 1995). This observation is very troublesome for theories that posit invariant constriction targets since the locations of the primary vocal tract constrictions vary significantly for the different /r/ configurations (e.g. the tongue body and tongue tip constrictions shown here). Instead, it appears that the only invariant target for American English /r/ is an acoustic or auditory perceptual target.

ning of movements toward these targets in an acoustic or auditory perceptual frame, and (2) the use of a directional mapping between the planning frame and the articulator frame to carry out the planned formant trajectories.

To understand this explanation, it is first important to note that simple target regions in acoustic or auditory perceptual space can correspond to complex regions in articulator space. The top half of Figure 12 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 the 7-dimensional articulator space of the Maeda articulator model (Maeda, 1990). This figure was produced by fixing five of the Maeda articulators at their neutral locations and varying the remaining two (tongue tip position and tongue body position) through their entire ranges to determine which configurations, after synthesis of a speech waveform based on the resulting area functions, produce formant frequencies that fall in the ranges specified in the top half of the figure. Note that the articulator space region is broken into two distinct sub-regions. The top sub-region roughly corresponds to a flattened tongue tip as in a retroflexed /r/, and the bottom sub-region roughly corresponds to a bunched tongue configuration as in a bunched /r/.

Figure 12 also includes arrows that indicate movement trajectories produced by the DIVA model in simulations reported in Guenther (1995b). The solid arrows indi-

cate the trajectory formed when moving from a /d/ configuration to an /r/ configuration, as in the word "drag". The solid arrow in the left half of the figure shows the movement trajectory as it is planned and carried out in formant space, and the solid arrow in the right half of the figure shows the articulator movements that were commanded by the model to realize this formant trajectory[6]. In this case, the model moves to the sub-region of articulatory space that corresponds to a retroflexed /r/. In other words, it uses a retroflexed configuration to produce /r/ when /r/ is preceded by /d/. The dashed arrows show the trajectories when /r/ is preceded by /g/. The planned trajectory in formant space moves to the same target region as when /r/ is preceded by /d/, but the corresponding articulator trajectory ends up in the sub-region of articulator space corresponding to bunched /r/ rather than retroflexed /r/. Relatedly, Espy-Wilson and Boyce (1994) describe a speaker who uses a bunched /r/ only when /r/ is adjacent to /g/ and a retroflexed /r/ for all of the other conditions in their experiment. The important thing to note about the model's explanation is that the directional mapping transforms planned formant trajectories, which go to a single target region in formant space, into articulator trajectories that can end up at *different* sub-regions in articulator space depending on phonetic context. Roughly speaking, the directional 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).

Figure 13 provides a second way to visualize this explanation. From a /g/ configuration (Figure 13a), the speech production system has learned that the articulator movements shown by the white arrows in the figure produce the formant changes needed to move toward the formant target for /r/. Carrying out these movements changes the formants, and the movements terminate when the formants have reached their target values. In this case, this occurs when the tongue reaches the bunched /r/ configuration schematized in Figure 13b. When starting from a /d/ configuration (Figure 13c), a different set of movements has been learned for carrying out the formant changes needed to reach the /r/ target (white arrows). Carrying out these movements leads to the retroflexed configuration schematized in Figure 13d, where again the invariant formant target region for /r/ is reached. 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 *changes* in articulator position by the directional map.

6. Strictly speaking, the target sub-regions for the tongue tip position and tongue body position shown in the right half of Figure 12 are valid only when the remaining five articulators are fixed in their neutral positions. As these other articulators are moved from their neutral positions by the model, the shapes of the target sub-regions for these two dimensions will change. This is because the articulator space sub-regions are actually seven-dimensional, and each sub-region plotted in Figure 12 represents only a single 2-D "slice" through the seven-dimensional sub-region. This approximation is valid for the explanation provided in the text, however, since the key point is simply that a single target region in acoustic space corresponds to two distinct subregions in articulator space.

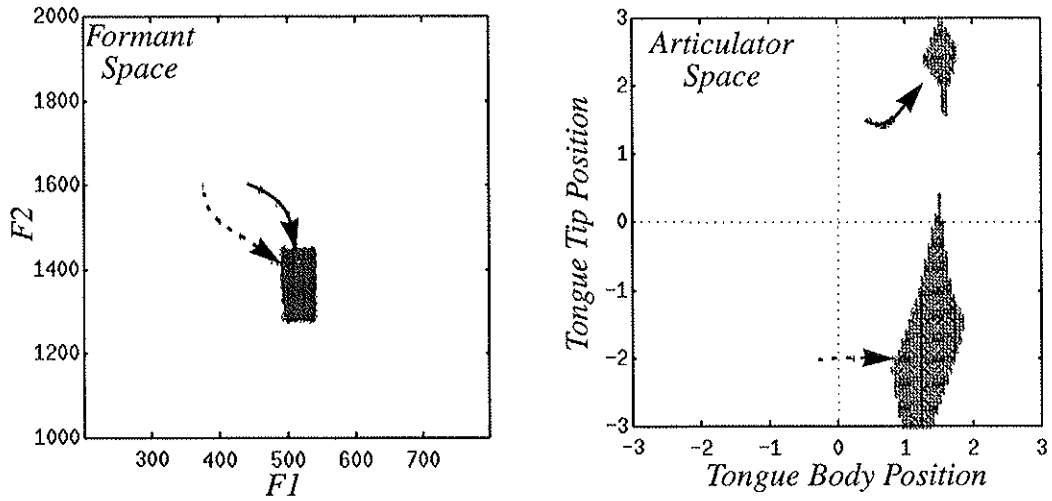

FIGURE 12. Relationship between a simple convex region corresponding to /r/ in acoustic space (left) and the corresponding region in articulator space (right). Arrows indicate model trajectories when producing /r/ starting from a /d/ configuration (solid lines) and from a /g/ configuration (dashed lines). Like many American English speakers, the model chooses a different configuration for /r/ depending on phonetic context.

Some caveats regarding this simulation should be made. 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 configuration and a bunched tongue configuration, are used to produce /r/ in different contexts), the model's configurations only very 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 as simulated did not include F3 in the planning space, this aspect is not captured here. The sublingual cavity that accompanies retroflex tongue shapes is likely to be partly responsible for lowering F3 when producing a retroflexed /r/ (Ken Stevens, personal communication). 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/ configurations that are more retroflexed. It should also be noted that the preceding phoneme appears to be only one of the factors that determines which configuration is used, so the explanation provided here is not a complete account of the variability seen in /r/ production. In spite of these caveats, the simulations reported here show how a model that utilizes a single target in acoustic or auditory perceptual space can use very different articulator configurations to reach this target if a directional mapping is used to transform desired acoustic/auditory changes into changes in articulator positions. Thus, such a model has the capacity to explain how different /r/ configurations can arise from a single invariant target, whereas a theory that posits invariant constriction targets does not.

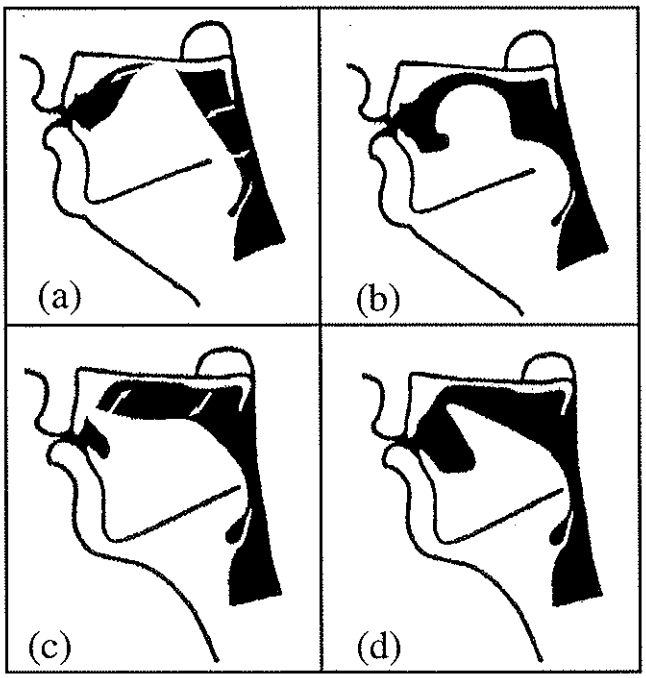

FIGURE 13. Schematic of the explanation for different /r/ configurations by the same speaker in two different contexts put forth by the DIVA model. (a) Vocal tract configuration used to produce /g/ in "grab". For this configuration, the motor system has learned that the tongue movements indicated by the white arrows in the figure can be used to produce the formant changes needed to reach the acoustic/auditory target for /r/. (b) Bunched /r/ configuration that would result from the movements shown in (a). (c) Schematized vocal tract configuration for producing /d/ in "drag", including the tongue movements learned by the motor system for changing the formants as needed to produce /r/ from the /d/ configuration (white arrows). (d) Retroflex /r/ configuration that would result from the movements shown in (c). [Portions of this figure were adapted from Delattre and Freeman (1968).]

Finally, the model's explanation provides answers to two more general questions concerning /r/ production. First, why is American English /r/ produced with very different configurations when this behavior is not typically seen for other sounds such as vowels? The answer provided here concerns the relationship between target regions in acoustic or auditory perceptual space and the corresponding regions in articulator space. Roughly speaking, the model predicts that, for vowels, the acoustic/auditory target region corresponds to a single articulator space region, but for /r/, two or more essentially distinct articulator space sub-regions correspond to the same acoustic/auditory target region[7] (see Figure 12). This prediction receives support from simulations of the DIVA model using the Maeda articulator set. The previous section showed how the "postural relaxation" component of the model leads to approximately invariant vocal tract shapes for vowels despite using no vocal tract shape targets. Interestingly, when the same model attempts to reach the acoustic space target for /r/ from different initial configurations, a bimodal distri-

bution of final configurations is found. This is illustrated in Figure 14. For some initial configurations a more bunched tongue shape is used, whereas a flatter tongue tip configuration is used for other initial configurations. These two different configurations correspond to the different articulator space sub-regions shown in Figure 12. It should be noted, however, that this simulation does not account for the low F3 characteristic of /r/ (due to the limitations on tongue retroflexion and sublingual cavity effects in the Maeda articulator set, as mentioned above), and it should thus be regarded as a preliminary result that only indicates the feasibility of the explanation put forth here, i.e., that two distinct sub-regions of articulator space may be used for some phonemes but not others. In particular, although one of the /r/ configurations produced by the model approximates a bunched /r/, the other configuration is at present a poor approximation to a retroflexed /r/.

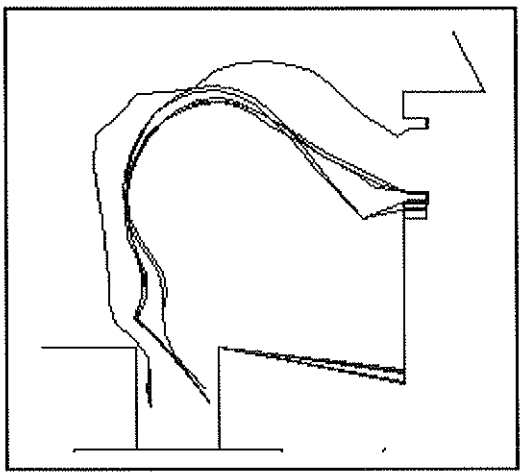

FIGURE 14. Bimodal distribution of configurations reached by the model when producing /r/ from different initial configurations. Unlike the vowel simulations reported in the previous section, which show an approximately invariant vocal tract shape for each vowel regardless of the vocal tract shape before the onset of the vowel, the model uses one of two different configurations when moving toward the auditory perceptual target for /r/. This behavior arises because two distinct sub-regions of articulator space are used to produce /r/, but only a single region of articulator space is used for each of the vowels simulated in Section 3.

The second question concerns the number of different vocal tract configurations used to produce /r/. I.e., are there two main types of /r/ articulation as suggested by

7. The actual situation is more complicated than outlined here. Having two distinct sub-regions in articulator space does not insure that the speech production mechanism will use both sub-regions. For example, one sub-region might correspond to a region of articulator space that is remote from the articulator configurations used to produce all other sounds and might therefore never be used. Furthermore, the region in articulator space corresponding to the acoustic/auditory target does not necessarily need to be composed of totally distinct subregions to lead to two or more configurations for the same sound. The stated prediction that more than one configuration will be used for a sound if and only if the articulator space region for that sound is broken into distinct sub-regions is thus only a first approximation, and additional factors will affect whether the model will use more than one configuration for a given sound in different contexts.

Espy-Wilson and Boyce (1994), three types as suggested by Hagiwara (1994, 1995), or an approximate continuum as suggested by Delattre and Freeman (1968) and Westbury et al. (1995)? The model's answer is that the number of articulations a speaker uses for /r/ corresponds to the number of essentially distinct sub-regions of articulator space that can be used by that speaker to produce the auditory cues corresponding to /r/. This number will vary from speaker to speaker since it depends on the exact shape of his/her vocal tract. This explanation provides a rationale for why some speakers use only one configuration while others use two or more. When looking across a range of speakers, one would expect an approximate continuum of /r/ articulations as suggested by the comprehensive study of Delattre and Freeman (1968) due to the approximate continuum of vocal tract shapes across speakers. The answers to these questions provided here, along with the effects of incorporating more realistic, speaker-specific vocal tract models into the DIVA framework, are being investigated in ongoing research.

6. Concluding remarks

The issue of the reference frame of target specification and movement planning is a crucial one for interpreting speech production data, designing new speech production experiments, and building a mechanistic understanding of speech production and the interplay between perception and production. This article has presented a four-part theoretical argument, supported by a range of experimental and modeling results, that favors theories postulating invariant auditory perceptual targets for the planning of speech movements for vowels and semivowels over theories postulating invariant constriction targets.

This theoretical treatment poses several clear and difficult challenges to proponents of constriction-based target theories. First, how can the central nervous system extract constriction locations and degrees from tactile and proprioceptive feedback, given that this mapping from orosensory feedback to a constriction representation is different for each individual, the mapping changes as the individual grows, and there is no direct feedback of constriction location and degree that can serve as a "teaching signal" for learning the mapping? Second, why are most subjects capable of compensating for a lip tube, in part by changing the locations and degrees of tongue body constrictions (Savariaux et al., 1995a,b), if they are utilizing invariant constriction targets? Third, why are trading relations between constrictions that are consistent with the idea of an invariant acoustic or auditory perceptual target seen during the production of several phonemes (Perkell et al., 1993, 1994)? Fourth, why are completely different constriction configurations often used by the same speaker for producing American English /r/?

Although we do not at present see how a constriction theory can account for these observations, we do not claim that it is impossible for a modified constriction theory to accomodate such results. Rather, we claim that no such theory can explain this collection of results in as simple a manner as an auditory perceptual target theory using a directional mapping between the planning frame and the articulator

frame, as implemented in the DIVA model of speech production. For example, a constriction target theory would have to posit two or more different targets for /r/ in some, but not all, speakers of American English, and furthermore would have to develop a rationale for why and how the speakers with multiple targets choose between them. The auditory perceptual target theory embodied by the DIVA model, however, provides an account of this result, including the cross-speaker variability, under the much simpler assumption of a single auditory perceptual target (see Section 5). The use of a directional mapping between the planning frame and the articulator frame in the DIVA model provides an account of how different vocal tract configurations for /r/ can automatically arise in different contexts without the need for additional machinery to determine which target to use on each occasion. A constriction target theory would have to explain how the nervous system can self-organize a useful representation of constriction locations and degrees from tactile and proprioceptive information in a very different reference frame, without direct constriction feedback to serve as a teaching signal. The job is much simpler for an auditory perceptual target theory since direct auditory perceptual feedback is available and can act as a teaching signal to train a forward model that transforms tactile and proprioceptive information into the auditory perceptual reference frame. The trading relations data of Perkell et al. (1993, 1994) and lip tube compensation data of Savariaux et al. (1995a,b) might be explained by hypothesizing an additional mechanism that somehow learns to trade between constrictions, but a complex theory of this sort starts to look very much like an auditory perceptual target theory in its simplest form.

In addition to posing challenges to constriction theories, we have accounted for the approximate invariance of constriction locations and degrees seen for vowels during normal speech. The explanation provided herein does not rely on invariant constriction targets in the production process, but instead utilizes invariant auditory perceptual targets and plans movement trajectories in auditory perceptual space. Approximate constriction invariance arises from a tendency to choose movements that keep the articulators near the centers of their movement ranges; this property is inherent to the neural mapping from the auditory perceptual planning space to the articulator movements used to carry out the planned trajectories. Simulation results verify that it is possible for a neural network to learn such a mapping during a babbling cycle, and that a controller using the mapping can account for approximate constriction invariance without explicit constriction targets for vowels (Section 3). Unlike invariant constriction target theories, this approach can also account for constriction variability when needed to overcome constraints on the articulators (Section 4), and the multiple vocal tract configurations for /r/ seen in many speakers of American English (Section 5).

We have been careful in this article to limit our strongest claims to vowels and semivowels. This is not done out of a conviction that consonant production is fundamentally different than vowel and semivowel production. In fact, we believe that many of the same arguments posed here may well hold for consonants, but we also feel that making a strong case for this claim will require further investigation. Other researchers have suggested that the different speech sound classes might be

produced in different ways (e.g., Fowler, 1980), including the suggestion that some sounds might use invariant acoustic/auditory targets or target regions while other sounds might use invariant vocal tract shape targets or target regions (e.g., Bailly, 1995; Perkell et al., 1997).

The model presented in this paper extends the convex region theory of the targets of speech to the domain of auditory perceptual targets. Guenther (1995a) describes how the convex region theory provides a unified account for many speech production phenomena, including motor equivalence, contextual variability, carryover coarticulation, anticipatory coarticulation, and variability due to changes in speaking rate. It is expected that most, if not all, of these explanations will also apply to target regions in auditory perceptual space. Related work by Guenther and Gjaja (1996) provides an account for the formation of auditory perceptual target regions, as evidenced by the perceptual magnet effect. These regions are hypothesized to arise as an emergent property of neural map formation in the auditory system.

References

[1] Abbs, J. H. (1986). Invariance and variability in speech production: A distinction between linguistic intent and its neuromotor implementation. In J. S. Perkell and D. H. Klatt (Eds.), Invariance and variability in speech processes (pp. 202-219). Hillsdale NJ: Erlbaum.

[2] Abbs, J. H., & Gracco, V. L. (1984). Control of complex motor gestures: Orofacial muscle responses to load perturbations of lip during speech. Journal of Neurophysiology, 51, 705-723.

[3] Baillieul, J., Hollerbach, J., & Brockett, R.W. (1984). Programming and control of kinematically redundant manipulators. Proceedings of the 23rd IEEE conference on decision and control (pp. 768-774). New York: IEEE.

[4] Bailly, G. (1995). Recovering place of articulation for occlusives in VCVs. Proceedings of the XIIIth International Conference of Phonetic Sciences (vol. 2, pp. 230-233). Stockholm, Sweden: KTH and Stockholm University.

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

[6] Bailly, G., Laboissière, R., and Galván, A. (1997). Learning to speak: Speech production and sensori-motor representations. In P. Morasso and V. Sanguineti (Eds.), Self-organization, computational maps and motor control (pp. 593-635). Amsterdam: North Holland.

[7] Borden, G. J. (1979). An interpretation of research on feedback interruption in speech. Brain and Language, 7, 307-319.

[8] Borden, G. J., Harris, K. S., & Oliver, W. (1973). Oral feedback I. Variability of the effect of nerve-block anesthesia upon speech. Journal of Phonetics, 1, 289-295.

[9] Brooks, V. B. (1986). The neural basis of motor control. New York: Oxford University Press.

[10] Browman, C., & Goldstein, L. (1990a). Tiers in articulatory phonology, with some implications for casual speech. In J. Kingston and M. E. Beckman (Eds.), Papers in laboratory phonology. I: Between the grammar and physics of speech (pp. 341-376). Cambridge, UK: Cambridge University Press.

[11] Browman, C., & Goldstein, L. (1990b). Gestural specification using dynamically-defined articulatory structures. Journal of Phonetics, 18, 299-320.

[12] Bullock, D., & Grossberg, S. (1988). Neural dynamics of planned arm movements: Emergent invariants and speed-accuracy properties during trajectory formation. Psychological Review, 95, 49-90.

[13] Bullock, D., Grossberg, S., & Guenther, F. H. (1993). A self-organizing neural network model for redundant sensory-motor control, motor equivalence, and tool use. Journal of Cognitive Neuroscience, 5, 408-435.

[14] Cameron, S. (1995). Self-organizing neural networks for visual navigation and adaptive control. Doctoral dissertation, Boston University, Boston.

[15] Coker, C. H. (1976). A model of articulatory dynamics and control. Proceedings of the IEEE, 64, 452-460.

[16] Cooper, S. (1953). Muscle spindles in the intrinsic muscles of the human tongue. Journal of Physiology, 122, 193-202.

[17] Cruse, H. (1986). Constraints for joint angle control of the human arm. Biological Cybernetics, 54, 125-132.

[18] Cruse, H., Brüwer, M., & Dean, J. (1993). Control of three- and four-joint arm movement: Strategies for a manipulator with redundant degrees of freedom. Journal of Motor Behavior, 25(3), 131-139.

[19] Delattre, P., & Freeman, D. C. (1968). A dialect study of American r's by x-ray motion picture. Linguistics, 44, 29-68.

[20] Desmurget, M., Prablanc, C., Rossetti, Y., Arzi, M., Paulignan, Y., Urquizar, C., & Mignot, J. (1995). Postural and synergic control for three-dimensional movements of reaching and grasping. Journal of Neurophysiology, 74, 905-910.

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

[22] Fant, G. (1992). Vocal tract area functions of Swedish vowels and a new three-parameter model. In J. Ohala, T. Neaty, B. Berwing, M. Hodge, and G. Wiebe (Eds.), ISCLP 92 Proceedings (vol. 1., pp. 807-810). Edmonton, Canada: University of Alberta.

[23] Flash, T. (1989). Generation of reaching movements: Plausibility and implications of the equilibrium trajectory hypothesis. Brain, Behavior, and Evolution, 33, 63-68.

[24] Fowler, C. A. (1980). Coarticulation and theories of extrinsic timing. Journal of Phonetics, 8, 113-133.

[25] Ghez, C., Gordon, J., & Ghilardi, M. F. (1995). Impairments of reaching movements in patients without proprioception. II. Effects of visual information on accuracy. Journal of Neurophysiology, 73, 361-372.

[26] Gomi, H., & Kawato, M. (1996). Equilibrium-point control hypothesis examined by measured arm stiffness during multijoint movement. Science, 272, 117-120.

[27] Gordon, J. & Ghez, C. (1991). Muscel receptors and spinal reflexes: The stretch reflex. In E. R. Kandel, J. H. Schwartz, & T. M. Jessell (Eds.), Principles of neural science, third edition (pp. 564-580). Norwalk, Connecticut: Appleton & Lange.

[28] Gordon, J., Ghilardi, M. F., & Ghez, C. (1995). Impairments of reaching movements in patients without proprioception. I. Spatial errors. Journal of Neurophysiology, 73, 347-360.

[29] Grajski, K. A., & Merzenich, M. M. (1990). Hebb-type dynamics is sufficient to account for the inverse magnification rule in cortical somatotopy. Neural Computation, 2, 71-84.

[30] Grossberg, S. (1976). Adaptive pattern classification and universal recoding: I. Parallel development and coding of neural feature detectors. Biological Cybernetics, 23, 121-134.

[31] Grossberg, S. (1980). How does a brain build a cognitive code? Psychological Review, 87, 1-51.

[32] Guenther, F. H. (1992). Neural models of adaptive sensory-motor control for flexible reaching and speaking. Doctoral dissertation, Boston University, Boston.

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

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

[35] 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.

[36] Guenther, F. H., & Gjaja, M. N. (1996). The perceptual magnet effect as an emergent property of neural map formation. Journal of the Acoustical Society of America, 100, 1111-1121.

[37] Guenther, F. H., & 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: North Holland.

[38] Hagiwara, R. (1994). Three types of American /r/. UCLA Working Papers in Phonetics, 88, 63-90.

[39] Hagiwara, R. (1995). Acoustic realizations of American /r/ as produced by women and men. UCLA Working Papers in Phonetics, 90.

[40] Hogan, N. (1984). An organizing principle for a class of voluntary movements. Journal of Neuroscience, 4, 2745-2754.

[41] Hoole,, P. (1987). Bite-block speech in the absence of oral sensibility. Proceedings of the XIIIth International Conference of Phonetic Sciences (vol. 4, pp. 16-19).

[42] Hore, J., Watts, S., & Vilis, T. (1992). Constraints on arm position when pointing in three dimensions: Donders' law and the Fick gimbal strategy. Journal of Neurophysiology, 68, 374-383.

[43] Iverson, P., and Kuhl, P.K. (1995). Mapping the perceptual magnet effect for speech using signal detection theory and multidimensional scaling. Journal of the Acoustical Society of America, 97, 553-562.

[44] Johnson, D., & Guenther, F. H. (1995). Acoustic space movement planning in a neural model of motor equivalent vowel production. World Congress on Neural Networks, Washington, D.C. (pp. 481-484). Mahwah, NJ: Lawrence Erlbaum Associates and INNS Press.

[45] Jordan, M. I. (1990). Motor learning and the degrees of freedom problem. In M. Jeannerod (Ed.), Attention and performance XIII (pp. 796-836). Hillsdale, NJ: Erlbaum.

[46] Klein, C. A., & Huang, C. (1983). Review of pseudoinverse control for use with kinematically redundant manipulators. IEEE Transactions on Systems, Man, and Cybernetics, SMC-13(2), 245-250.

[47] Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. Biological Cybernetics, 43, 59-69.

[48] Kröger, B. J. (1993). A gestural production model and its application to reduction in German. Phonetica, 50, 213-233.

[49] Kuhl, P.K. (1991). Human adults and human infants show a 'perceptual magnet effect' for the prototypes of speech categories, monkeys do not. Perception & Psychophysics, 50, 93-107.

[50] Kuhl, P.K. (1995). Mechanisms of developmental change in speech and language. In Elenius, K., and Branderud, P. (eds.): Proceedings of the XIIIth International Congress of Phonetic Sciences (vol. 2, pp. 132-139). Stockholm: KTH and Stockholm University.

[51] Laboissière, R., & Galvan, A. (1995). Inferring the commands of an articulatory model from acoustical specifications of stop/vowel sequences. Proceedings of the XIIIth International Conference of Phonetic Sciences (vol. 1, pp. 358-361). Stockholm, Sweden: KTH and Stockholm University.

[52] Laboissière, R., Ostry, D. J., & Perrier, P. (1995). A model of human jaw and hyoid motion and its implications for speech production. Proceedings of the XIIIth International Conference of Phonetic Sciences (vol. 2, pp. 60-67). Stockholm, Sweden: KTH and Stockholm University.

[53] Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. Psychological Review, 74, 431-461.

[54] Liberman, A. M., & Mattingly, I. G. (1985). The motor theory of speech revisited. Cognition, 21, 1-36.

[55] Lieberman, P., & Blumstein, S. E. (1988). Speech physiology, speech perception, and acoustic phonetics. Cambridge, UK: Cambridge University Press.

[56] Liégeois, A. (1977). Automatic supervisory control of the configuration and behavior of multibody mechanisms. IEEE Transactions on Systems, Man, and Cybernetics, SMC-7(12), 868-871.

[57] Lindblom, B., Lubker, J., & McAllister, R. (1977). Compensatory articulation and the modeling of normal speech production behavior. In R. Carré, R. Descout, and M. Wajskop (Eds.), Articulatory modeling and phonetics (Proceedings from Symposium at Grenoble, G.A.L.F.)

[58] 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, 7, 147-161.

[59] MacNeilage, P. F., Rootes, T. P., & Chase, R. A. (1967). Speech production and perception in a patient with severe impairment of somesthetic perception and motor control. Journal of Speech and Hearing Research, 10, 449-467.

[60] Maeda, S. (1990). Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal tract shapes using an articulatory model. In

W.J. Hardcastle and A. Marchal (Eds.), Speech production and speech modelling (pp. 131-149). Boston: Kluwer Academic Publishers.

[61] Matthews, P. B. C. (1972). Mammalian muscle receptors and their central actions. London: Edward Arnold.

[62] Mermelstein, P. (1973). Articulatory model for the study of speech production. Journal of the Acoustical Society of America, 53, 1070-1082.

[63] Miller, J. D. (1989). Auditory-perceptual interpretation of the vowel. Journal of the Acoustical Society of America, 85, 2114-2134.

[64] Miller, L.E., Theeuwen, M., & Gielen, C.C. (1992). The control of arm pointing movements in three dimensions. Experimental Brain Research, 90, 415-426.

[65] Morasso, P. (1981). Spatial control of arm movements. Experimental Brain Research, 42, 223-227.

[66] Mussa-Ivaldi, F. A., & Hogan, N. (1991). Integrable solutions of kinematic redundancy via impedance control. International Journal of Robotics Research, 10, 481-491.

[67] Narayanan, S., Alwan, A., & Haker, K. (1995). An articulatory study of liquid approximants in American English. Proceedings of the XIIIth International Conference of Phonetic Sciences (vol. 3, pp. 576-579). Stockholm, Sweden: KTH and Stockholm University.

[68] Ong, D., & Stone, M. (1997). Three-dimensional vocal tract shapes in /r/ and /l/: A study of MRI, ultrasound, electropalatography, and acoustics. Phonoscope, in press.

[69] Ostry, D. J., Gribble, P. L., and Gracco, V. L. (1996). Coarticulation of jaw movements in speech production: Is context sensitivity in speech kinematics centrally planned? Journal of Neuroscience, 16, 1570-1579.

[70] Perkell, J. S. (1980). Phonetic features and the physiology of speech production. In B. Butterworth (Ed.), Language production, volume 1: Speech and talk (pp. 337-372). New York: Academic Press.

[71] Perkell, J. S. (1997). Articulatory processes. In W. J. Hardcastle and J. Laver (Eds.), The handbook of phonetic sciences (pp. 333-370). Cambridge, MA: Blackwell Publishers.

[72] Perkell, J. S., Matthies, M. L., Lane, H., Wilhelms-Tricarico, R., Wozniak, J., & Guiod, P. (1997). Speech motor control: Segmental goals and the use of feedback. Submitted to Speech Communication.

[73] Perkell, J. S., and Klatt, D. H. (1986). Invariance and variability in speech processes. Hillsdale NJ: Erlbaum.

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

[75] Perkell, J. S., Matthies, M. L., Svirsky, M. A., & 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.

[76] Poggio, T., & Girosi, F. (1989). A theory of networks for approximation and learning. AI Memo No. 1140, Massachusetts Institute of Technology.

[77] Rabiner, L. R., & Schafer, R. W. (1978). Digital processing of speech signals. Englewood Cliffs, NJ: Prentice-Hall.

[78] Rosenbaum, D. A., Engelbrecht, S. E., Bushe, M. M., & Loukopoulos, L. D. (1993). Knowledge model for selecting and producing reaching movements. Journal of Motor Behavior, 25, 217-227.

[79] Rosenbaum, D. A., Loukopoulos, L. D., Meulenbroek, R. G. J., Vaughan, J., & Engelbrecht, S. E. (1995). Planning reaches by evaluating stored postures. Psychological Review, 192, 28-67.

[80] Rubin, P., Baer, T., & Mermelstein, P. (1981). An articulatory synthesizer for perceptual research. Journal of the Acoustical Society of America, 70, 321-328.

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

[82] Savariaux, C. (1995). Étude de l'espace de contrôle distal en production de la parole: Les enseignements d'une perturbation à l'aide d'un tube labial. Doctoral dissertation, l'Institut National Polytechnique de Grenoble, Grenoble, France.

[83] Savariaux, C., Perrier, P., & Orliaguet, J. P. (1995a). Compensation strategies for the perturbation of the rounded vowel [u] using a lip tube: A study of the control space in speech production. Journal of the Acoustical Society of America, 98, 2428-2442.

[84] Savariaux, C., Perrier, P., & Schwartz, J. L. (1995b). Perceptual analysis of compensatory strategies in the production of the French rounded vowel [u] perturbed by a lip tube. Proceedings of the XIIIth International Congress of Phonetic Sciences (vol. 3, pp. 584-587). Stockholm, Sweden: KTH and Stockholm University.

[85] Soechting, J.F., Buneo, C.A., Herrmann, U., & Flanders, M. (1995). Moving effortlessly in three dimensions: Does Donders' law apply to arm movement? Journal of Neuroscience, 15, 6271-6280.

[86] Stevens, K. N., & Perkell, J. S. (1977). Speech physiology and phonetic features. In M. Sawashima and F. S. Cooper (Eds.), Dynamic aspects of speech production: Current results, emerging problems, and new instrumentation (pp. 323-341). Tokyo: University of Tokyo Press.

[87] Stokbro, K., Umberger, D. K., & Hertz, J. A. (1990). Exploiting neurons with localized receptive fields to learn chaos. Complex Systems, 4, 603-622.

[88] Stone, M. (1991). Toward a model of three-dimensional tongue movement. Journal of Phonetics, 19, 309-320.

[89] Sutton, G. G. III, Reggia, J. A., Armentrout, S. L., & D'Autrechy, C. L. (1994). Cortical map reorganization as a competitive process. Neural Computation, 6, 1-13.

[90] Uno, Y., Kawato, M., & Suzuki, R. (1989). Formation and control of optimal trajectory in human multijoint arm movement. Biological Cybernetics, 61, 89-101.

[91] von der Malsburg, C. (1973). Self-organization of orientation sensitive cells in the striata cortex. Kybernetik, 14, 85-100.

[92] Westbury, J. R., Hashi, M., & Lindstrom, M. J. (1995). Differences among speakers in articulation of American English /r/: An x-ray microbeam study. Proceedings of the XIIIth International Conference of Phonetic Sciences (vol. 4, pp. 50-57). Stockholm, Sweden: KTH and Stockholm University.

[93] Wilhelms-Tricarico, R. (1995). Physiological modeling of speech production: Methods for modeling soft-tissue articulators. Journal of the Acoustical Society of America, 97, 3085-3098.

[94] Wilhelms-Tricarico, R. (1996). A biomechanical and physiologically-based vocal tract model and its control. Journal of Phonetics, 24, 23-38.

[95] Wolpert, D.M., Ghahramani, Z., & Jordan, M.I. (1994). Perceptual distortion contributes to the curvature of human reaching movements. Experimental Brain Research, 98, 153-156.

[96] Wolpert, D.M., Ghahramani, Z., & Jordan, M.I. (1995). Are arm trajectories planned in kinematic or dynamic coordinates? An adaptation study. Experimental Brain Research, 103, 460-470.

Appendix A: Derivation of radial basis function learning rules.

The directional mapping between auditory perceptual space and articulator space is learned using a form of hyperplane radial basis function (HRBF) network (Cameron, 1995; see also Poggio and Girosi, 1989; Stokbro, Umberger, and Hertz, 1990).

A three-dimensional auditory space, x, is mapped to a seven-dimensional articulator space, θ, using the following discrete-time approximation to equation 7:

$$\Delta\theta = G(\theta)\Delta x + R(\theta)$$

where $G(\theta)$ is a three by seven matrix and $R(\theta)$ is a seven-dimensional vector.

The matrix $G(\theta)$:

Each entry of $G(\theta)$ is represented by an HRBF network $g_{ij}(\theta)$. Each hyperplane basis function has one weight, v, to indicate the magnitude of the data under its receptive field, and a set of weights, w, which allow it to linearly approximate the slope of the data under its receptive field. The output of the network $g_{ij}(\theta)$ is given by:

$$g_{ij}(\theta) = \sum_k \left(\frac{A_{ijk}}{\sum_k A_{ijk}} \right) \left(v_{ijk} + \sum_l c_{ijkl} w_{ijkl} \right)$$

where k is the index of the basis function, the vector c_{ijk} is a measure of the distance between the input value θ and the center of the k^{th} basis function in that network, and A_{ijk} is the activation of the basis function (drops off in a Gaussian fashion from the center).

$$c_{ijkl} = \frac{x_l - \mu_{ijkl}}{\sigma_{ijkl}} \qquad A_{ijk} = \exp\left[-\sum c_{ijkl}^2 \right]$$

Here μ_{ijkl} and σ_{ijkl} are, respectively, the center and standard deviation along the l^{th} dimension of the k^{th} Gaussian activation function (in the ij^{th} network). For

simpler notation, h_{ijk} will refer to the normalized activation of the k^{th} basis function:

$$h_{ijk} = \frac{A_{ijk}}{\sum_k A_{ijk}}$$

The weights v_{ijk} and w_{ijkl} are updated by gradient descent according to the equations:

$$\Delta v_{ijk} = -\alpha\left(\frac{\partial H_1}{\partial v_{ijk}}\right)$$

and

$$\Delta w_{ijkl} = -\alpha\left(\frac{\partial H_1}{\partial w_{ijkl}}\right)$$

where $H_1 = \sum_i (\Delta\theta_{Bi} - \Delta\theta_i)^2$ and α is the learning rate.

Applying the chain rule yields:

$$\begin{aligned}\Delta v_{ijk} &= -2\alpha(\Delta\theta_{Bi} - \Delta\theta_i)\left(\frac{\partial(\Delta\theta_i)}{\partial g_{ij}}\right)\left(\frac{\partial g_{ij}}{\partial v_{ijk}}\right) \\ &= -2\alpha(\Delta\theta_{Bi} - \Delta\theta_i)(\Delta x_j)(h_{ijk})\end{aligned}$$

$$\begin{aligned}\Delta w_{ijkl} &= -2\alpha(\Delta\theta_{Bi} - \Delta\theta_i)\left(\frac{\partial(\Delta\theta_i)}{\partial g_{ij}}\right)\left(\frac{\partial g_{ij}}{\partial w_{ijkl}}\right) \\ &= -2\alpha(\Delta\theta_{Bi} - \Delta\theta_i)(\Delta x_j)(h_{ijk}c_{ijkl})\end{aligned}$$

The vector R(θ):

Each entry of R(θ) is also represented by an HRBF network, $r_i(\theta)$, with output given by:

$$r_i(\theta) = \sum_k \left(\frac{A_{ik}}{\sum_k A_{ik}}\right)\left(v_{ik} + \sum_l c_{ikl} w_{ikl}\right).$$

The weights for this network are updated to minimize the second cost function:

$$H_2 = \sum_i (\Delta\theta_{Bi} - \Delta\theta_i)^2 + \beta_i \left(\frac{\theta_i - \theta_i^c}{\theta_i^r} \right)^2$$

so

$$\Delta v_{ik} = -\alpha \left(\frac{\partial H_2}{\partial v_{ik}} \right) = -\alpha \left(\left(\frac{\partial H_2}{\partial (\Delta\theta_i)} \right) \left(\frac{\partial (\Delta\theta_i)}{\partial r_i} \right) + \left(\frac{\partial H_2}{\partial \theta_i} \right) \left(\frac{\partial \theta_i}{\partial r_i} \right) \right) \left(\frac{\partial r_i}{\partial v_{ik}} \right)$$

and

$$\Delta w_{ikl} = -\alpha \left(\frac{\partial H_2}{\partial w_{ikl}} \right) = -\alpha \left(\left(\frac{\partial H_2}{\partial (\Delta\theta_i)} \right) \left(\frac{\partial (\Delta\theta_i)}{\partial r_i} \right) + \left(\frac{\partial H_2}{\partial \theta_i} \right) \left(\frac{\partial \theta_i}{\partial r_i} \right) \right) \left(\frac{\partial r_i}{\partial w_{ikl}} \right).$$

What is $\frac{\partial \theta_i}{\partial r_i}$? Considering that this learning process is implemented in a discrete time system, what we are really interested in is $\frac{\partial \theta_i(t)}{\partial r_i(t)}$ for time step t. Assuming $\theta_i(t) = \Delta\theta_i(t) + \theta_i(t-1)$ then $\frac{\partial \theta_i(t)}{\partial r_i(t)} = \frac{\partial (\Delta\theta_i(t))}{\partial r_i(t)} + \frac{\partial \theta_i(t-1)}{\partial r_i(t)}$. The right-most term is zero for a causal system, so we set $\frac{\partial \theta_i}{\partial r_i} = \frac{\partial (\Delta\theta_i)}{\partial r_i}$. Substituting this into our derivation we have:

$$\Delta v_{ik} = -\alpha \left[2(\Delta\theta_{Bi} - \Delta\theta_i) + \frac{2\beta_i}{\theta_i^r} \left(\frac{\theta_i - \theta_i^c}{\theta_i^r} \right) \right] \left(\frac{\partial (\Delta\theta_i)}{\partial r_i} \right) \left(\frac{\partial r_i}{\partial v_{ik}} \right)$$

$$= -2\alpha \left[(\Delta\theta_{Bi} - \Delta\theta_i) + \frac{\beta_i}{\theta_i^r} \left(\frac{\theta_i - \theta_i^c}{\theta_i^r} \right) \right] h_{ik}$$

$$\Delta w_{ikl} = -\alpha\left[2(\Delta\theta_{Bi} - \Delta\theta_i) + \frac{2\beta_i}{\theta_i^r}\left(\frac{\theta_i - \theta_i^c}{\theta_i^r}\right)\right]\left(\frac{\partial(\Delta\theta_i)}{\partial r_i}\right)\left(\frac{\partial r_i}{\partial w_{ikl}}\right)$$

$$= -2\alpha\left[(\Delta\theta_{Bi} - \Delta\theta_i) + \frac{\beta_i}{\theta_i^r}\left(\frac{\theta_i - \theta_i^c}{\theta_i^r}\right)\right]h_{ik}c_{ikl}\,.$$