1994-04

Speech Sound Acquisition, Coarticulation, and Rate Effects in a Neural Network Model of Speech Production

SPEECH SOUND ACQUISITION, COARTICULATION, AND RATE EFFECTS IN A NEURAL NETWORK MODEL OF SPEECH PRODUCTION

Frank H. Guenther

April 1994
Revised: October 1994
Revised: January 1995

Technical Report CAS/CNS-94-012

Boston University Center for Adaptive Systems and
Department of Cognitive and Neural Systems
111 Cummington Street
Boston, MA 02215

Speech Sound Acquisition, Coarticulation, and Rate Effects in a Neural Network Model of Speech Production

Running title: Acquisition, Coarticulation, and Rate Effects

Frank H. Guenther*

Boston University
Center for Adaptive Systems and
Department of Cognitive and Neural Systems
111 Cummington Street
Boston, MA, 02215
Fax Number: (617) 353-7755

Psychological Review, in press

*Supported in part by AFOSR F49620-92-J-0499. The author would like to thank Dan Bullock, Elliot Saltzman, and two anonymous reviewers for their insightful suggestions on an earlier draft of this paper.

ABSTRACT

This article describes a neural network model of speech motor skill acquisition and speech production that explains a wide range of data on contextual variability, motor equivalence, coarticulation, and speaking rate effects. Model parameters are learned during a babbling phase. To explain how infants learn phoneme-specific and language-specific limits on acceptable articulatory variability, the learned speech sound targets take the form of multi-dimensional convex regions in orosensory coordinates. Reduction of target size for better accuracy during slower speech (in the spirit of the speed-accuracy trade-off described by Fitts' law) leads to differential effects for vowels and consonants, as seen in speaking rate experiments that have been previously taken as evidence for separate control processes for the two sound types. An account of anticipatory coarticulation is posited wherein the target for a speech sound is reduced in size based on context to provide a more efficient sequence of articulator movements. This explanation generalizes the well-known look-ahead model of coarticulation to incorporate convex region targets. Computer simulations verify the model's properties, including linear velocity/distance relationships, motor equivalence, speaking rate effects, and carryover and anticipatory coarticulation.

1. Introduction

The primary goal of the modeling work described in this article is to provide a coherent theoretical framework that provides explanations for a wide range of data concerning the articulator movements used by humans to produce speech sounds. This is carried out by formulating a model that transforms strings of phonemes into continuous articulator movements for producing these phonemes. This study of speech production is largely motivated by the following question of speech acquisition: *How does an infant acquire the motor skills needed to produce the speech sounds of his/her language?* Speech production involves complex interactions between several different reference frames. A phonetic frame describes the sounds a speaker wishes to produce, and the signals that convey these sound units to a listener exist within an acoustic frame. Tactile and proprioceptive signals form an orosensory frame (e.g., Perkell, 1980) that describes the shape of the vocal tract, and the muscles controlling the positions of individual articulators make up an articulatory frame. The parameters governing the interactions between these frames cannot be fixed at birth. One reason for this is the language specificity of these interactions. For example, English listeners distinguish between the sounds /r/ and /l/, but Japanese listeners do not. Corresponding differences are seen in the articulator movements of the two groups (Miyawaki et al., 1975). Thus, despite some obvious commonalities between the phonetics of different languages (e.g. widespread use of consonants like /d/, /n/, and /s/ across the world's languages), the precise nature of mappings between acoustic goals and articulator movements depends on the language being spoken. Interactions between reference frames must also be time-varying. As an infant grows, physical characteristics such as the length of the vocal tract and the shapes of articulators change. Temporary or permanent damage to the articulators may also occur. Such changes will affect the acoustic signal that is produced with a given set of motor commands. Maintaining the ability to properly produce important acoustic features thus requires that parameters governing the mappings between phonetic, acoustic, orosensory, and motor frames change with time.

Two important goals motivate the design of the present model. First, the resulting model should be *computational*; that is, it should be described in sufficient mathematical detail that its properties can be verified through computer simulation. The speech production mechanism is responsible for amazingly fast, flexible, and efficient movements. For example, speech production is inherently motor equivalent: many different motor actions can be

used to produce the same speech sound. A speaker may speak normally, using upward and downward movements of the jaw, or he/she can speak with the jaw clenched on a pipe. Production of a given speech sound in these two cases requires a completely different set of articulator positions and movements, yet humans automatically compensate for such constraints (e.g., Abbs and Gracco, 1984; Folkins and Abbs, 1975; Kelso, Tuller, Vatikiotis-Bateson, and Fowler, 1984; Lindblom, Lubker, and Gay, 1979). Furthermore, coarticulation greatly increases the efficiency of articulator movements. A model of speech motor skills should embody these competencies. However, as the complexity of a model increases to cover wider ranges of data, verification of the model's properties becomes increasingly difficult. Computer simulation becomes very desireable, if not mandatory, for verifying performance. The speech production literature contains very few examples of such computational models, but some very important contributions have been made. The dynamic articulatory model of Henke (1966) represented the first use of computer technology to generate complex movements of model articulators. Central concepts of this model such as the look-ahead model of coarticulation are still actively discussed in the speech production literature (e.g., Boyce, Krakow, Bell-Berti, and Gelfer, 1990; Wood, 1991). More recently, Saltzman and Munhall (1989) describe the most complete computational model of speech production to date. This impressive model, called the *task-dynamic* model, has been used to explain a wide range of coarticulation and motor equivalence data (see also the related work of Kröger, 1993).

The second goal is that the model should be *self-organizing*; that is, its parameters should be tuned based only on information available to an infant. The precise nature of the mappings between reference frames required for speech are language-specific and depend on things that change with time such as the lengths of the articulators and the strengths of the muscles. Thus, the human speech production system must adaptively organize appropriate mappings. The models mentioned above do not deal with the problem of adaptive organization of model parameters. Instead, appropriate parameter values were hand crafted by the modelers. In fact, MacNeilage and Davis (1990) lament that "there is at present no unified view of how [speech] motor control develops" due to the lack of attention to speech acquisition in the speech production literature (p. 454). In infants, babbling comprises an action-perception cycle that can be used to tune the parameters of the production system; the current model uses such a babbling cycle to learn mappings between reference frames. Other recent adaptive models have been posited for learning the relationship between

muscle EMG and articulator movements (Hirayama, Vatikiotis-Bateson, Kawato, and Jordan, 1992) and for use in speech synthesis using a model of the speech articulators (Bailly, Laboissière, and Schwartz, 1991).

To achieve these goals, the current model is formulated as an adaptive neural network. Two mappings are learned during babbling: (1) a *phonetic-to-orosensory* mapping wherein acceptable ranges of orosensory variables are learned for each speech sound, and (2) an *orosensory-to-articulatory* mapping wherein desired movements in orosensory space are mapped into articulator motor commands. The model is called DIVA after this latter mapping from **D**irections (in orosensory space) **I**nto **V**elocities of **A**rticulators, and has been briefly introduced in Guenther (1992; 1994). The learning processes use only information available to an infant (i.e., there are no "training sets" for the system's mappings as in standard backpropagation algorithms), and all learning laws governing the model's connections"synapses" use only information directly available from the pre- and post-synaptic "cells".

The answer embodied by the DIVA model to the question posed in the opening paragraph leads to a major theme of this article: *Insights gained from the study of speaking skill acquisition lead to novel and elegant explanations for long-studied speech production phenomena including motor equivalence, motor variability, speaking rate effects, and coarticulation.* This can be seen by looking at the forms of the two mappings learned by the model.

The phonetic-to-orosensory mapping specifies a vocal tract target for each speech sound. To explain how infants learn phoneme-specific and language-specific limits on acceptable articulatory variability, the targets take the form of *convex regions* in orosensory coordinates defining the shape of the vocal tract. A convex region is a multidimensional region such that for any two points in the region, all points on a line segment connecting these two points are also in the region. Two examples of convex regions are schematized in Figure 1. It is these regions, rather than specific configurations, that act as the vocal tract targets[1]. Convex region targets lead directly to explanations of motor variability and carryover coarticulation. Furthermore, shrinking of the target region for better accuracy dur-

1. From a dynamical systems viewpoint, this corresponds to using convex region attractors rather than point attractors.

ing slower speech (as suggested by the well-known speed-accuracy trade-off for movement control; e.g., Woodworth, 1899; Fitts, 1954) leads to differential effects for vowels and consonants: the speed of vowel movements remains approximately constant or even increases, whereas the speed of consonant movements decreases. This is in concert with experimental data on speaking rate effects (e.g., Gay, Ushijima, Hirose, and Cooper, 1974) that were previously taken as evidence for separate control structures for vowels and consonants (e.g., Fowler, 1980). The current work shows how a single control process can lead to these differential effects, with the effects arising due to inherent differences in the shapes of the target convex regions for vowels and consonants. The convex region theory also leads to an explanation of anticipatory coarticulation wherein the target region for a speech sound is reduced in size based on context in order to provide a more efficient sequence of articulator movements.

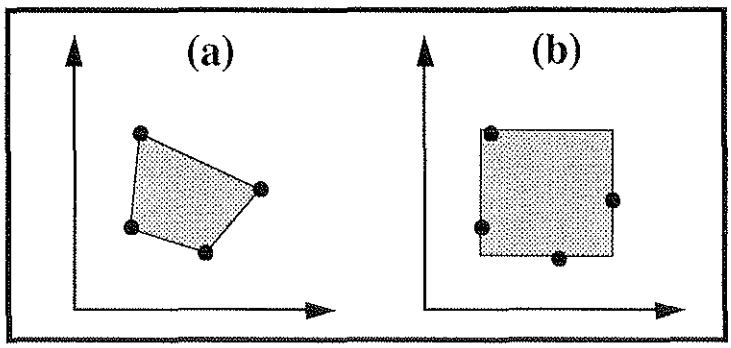

FIGURE 1. Two examples of convex regions. A convex region is a region such that for any two points in the region, all points on a line segment connecting the two points are also in the region. For a given set of points, a *convex hull* is the minimal convex region that encompasses these points. This is schematized for four points in (a). The convex regions for the speech sound targets as learned in the present implementation of the model are schematized in (b). These regions are defined by independent ranges along each dimension. This form of convex region is used to simplify learning and performance in the neural network described herein.

The orosensory-to-articulatory mapping transforms orosensory targets into appropriate articulator movements. An appropriate mapping from vocal tract targets to articulator movements is required to achieve automatic compensation for unexpected or unusual conditions such as a bite block or a perturbed articulator. In the task-dynamic model of Saltzman and Munhall (1989), this is accomplished through a complex dynamical system. The complexity of this dynamical system is largely due to the redundant nature of the mapping between vocal tract configurations and articulator positions; that is, many different combinations of articulator positions can be used to produce a single vocal tract configuration.

The DIVA model uses a much simpler redundant mapping between desired directions of movement in vocal tract configuration space and velocities of the articulators. The direction-to-velocity nature of this mapping not only results in motor equivalence, but also makes learning of the orosensory-to-articulatory mapping much simpler (Bullock, Grossberg, and Guenther, 1993 and Guenther, 1992 for a related discussion concerning the learning and use of a direction-to-velocity mapping to generate motor equivalent arm movements). This mapping leads to a well-known property of human speech articulator control: articulator velocities are directly related to movement distance (see Section 5 below). Investigation of the orosensory-to-articulatory mapping also reveals that articulators automatically organize into task-specific groupings or *coordinative structures* (Easton, 1972; see Section 3.1 below) during the learning process. Coordinative structures have long been hypothesized to play an important role in efficient movement control (Fowler, 1980; Saltzman and Kelso, 1987) and have also been observed in experimental data (e.g., Kelso et al., 1984).

Before proceeding to the model description, it should be noted that although this study of articulatory phonetics necessarily touches on many important unresolved issues in linguistics and phonology, the model addresses these issues only when they are directly relevant to the articulation of a string of sounds as specified by higher-level brain centers. For example, no attempt is made here to explain why humans do not produce arbitrary phoneme strings but instead apparently follow certain rules that determine which sounds can be produced in sequence; it is simply assumed that only appropriately structured strings will be sent to the modeled speech production mechanism. Likewise, many issues concerning the development of speech and language in children are touched upon but not directly addressed. Instead, attention is paid only to those aspects of infant development relevant to the acquisition of the motor skills necessary for the production of speech sounds independent of any underlying linguistic meaning or syllabic structure. In those instances where the model comes in contact with such issues, the assumptions concerning linguistics, phonology, or development will typically be as loose and general as possible. For example, the model is capable of producing arbitrary phoneme strings even though human speakers cannot. Because of their generality, it is hoped that these assumptions will remain valid when the related linguistic and developmental issues are resolved.

2. Overview of the DIVA Model

A block diagram of the DIVA model is shown in Figure 2. The model uses two different kinds of neural structure to represent information: vectors and maps. A *vector* is a set of antagonistic cell pairs that each code a different dimension in the space being represented (i.e., the input space); the pattern of activity across these cells codes the current position in this space. The notation "+" will be used to index a cell in an antagonistic pair whose activity increases for increasing values along the corresponding dimension of the input space, and "-" will be used to index the cell whose activity decreases for increasing values along the corresponding dimension of input space. This kind of push-pull coding is useful when both positive and negative displacements along a dimension need to be represented by a positive activity. For example, we will see below how the Orosensory Direction Vector codes desired movements of the vocal tract. Only positive activity of Orosensory Direction Vector cells can cause movements of the articulators, so it is necessary to represent both desired increases in position and desired decreases in position with positive activity of some cell in the Orosensory Direction Vector. Therefore, antagonistic pairs are needed to code desired movements in this vector. A *map* is a set of cells wherein each cell codes a small region in the input space. Only one cell can be maximally active in a map, and this cell alone codes the current position in the input space. Antagonistic cell pairing, vector representations, and map representations have been widely reported in the neurophysiological literature (e.g., Grobstein, 1991; Penfield and Rasmussen, 1950; Sakata, Shibutani, and Kawano, 1980).

The DIVA model incorporates information from four distinct reference frames: an acoustic frame, a phonetic frame, an orosensory (somatosensory) frame, and an articulatory (motor) frame. Signals in an acoustic frame make up the medium through which speech is communicated; the true job of the speech production mechanism is the creation of an appropriate set of acoustic signals to convey linguistic units from the speaker to listeners. Transduction and processing of these acoustic signals by the auditory system results in a phonetic reference frame. The phonetic frame in DIVA consists of the set of speech sounds that the model learns to produce. Signals from tactile and proprioceptive receptors form an orosensory frame that provides information about the shape of the vocal tract, which determines the sounds being produced. Evidence for a key role for orosensory information in normal speech production includes the inability of individuals with deficits

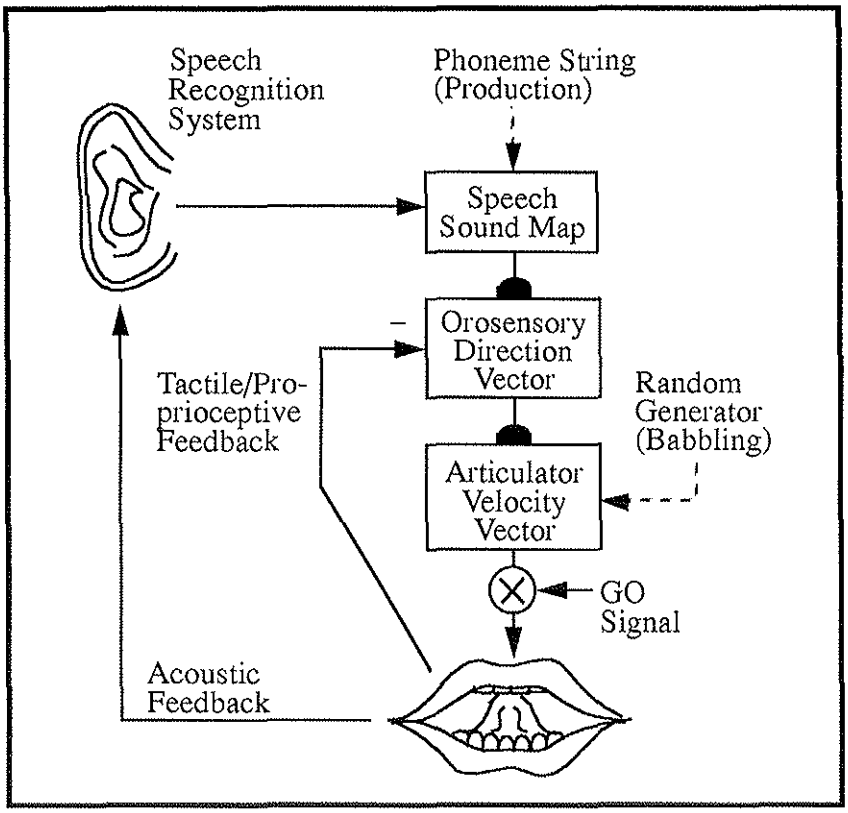

FIGURE 2. Overview of the DIVA model. Learned mappings are indicated by filled semicircles.

in somesthetic perception but no auditory or motor system damage to produce intelligible speech (e.g., MacNeilage, Rootes, and Chase, 1967) and the inability for subjects to properly compensate for a bite block when tactile information is disrupted (Lindblom, Lubker, and MacAllister, 1977; Lindblom et al., 1979). It should be noted, however, that other studies suggest that degraded but intelligible speech can sometimes be produced when somatosensory feedback from the vocal tract is interrupted (see Borden, 1979), suggesting a role for efference copies of commanded articulator movements in controlling speech. It is thus expected that at least an approximate representation of vocal tract shape within the orosensory frame described here can be formed from efference copies of motor outflow comands in addition to tactile and proprioceptive feedback from the vocal tract. Finally, an articulatory (motor) reference frame describes the commands to individual articulators and muscles to produce the movements that result in speech.

There are two learned mappings between these reference frames (shown as filled semicircles in Figure 2): a phonetic-to-orosensory mapping, and an orosensory-to-articulatory mapping. The parameters of these mappings are tuned during the babbling phase described below. A third, acoustic-to-phonetic mapping is approximated in the model by the Speech Recognition System as described below.

Finally, there are two forms of feedback in DIVA. Acoustic feedback is used for acquiring the orosensory targets corresponding to speech sounds, and orosensory feedback is used for both acquisition of speaking skills and for normal speech production.

Simulations of the DIVA model incorporate a babbling phase, during which the learned mappings are tuned, and a performance phase, during which the model produces phoneme strings specified by the modeler. These phases will now be briefly described, followed by descriptions of the various model components shown in Figure 2.

2.1. The Babbling Phase

Babbling during an infant's first year of life is made up of several overlapping stages (e.g., Oller, 1980; Stark, 1980). In the first two months, infants pass through a *phonation stage* (Oller, 1980) wherein speech-like sounds are relatively rare. The few speech-like sounds that are seen at this stage consist largely of phonation with the mouth closed or nearly closed. This is followed by a *goo stage* (2 to 3 months of age) wherein infants begin to produce very crude syllable-like sequences consisting largely of velar consonant-like elements in combination with the vowel-like elements seen during the phonation stage. At about 4 to 6 months of age, most infants enter the *expansion stage*, characterized by the emergence of several new sound types, including bilabial and labiolingual trills ("raspberries"), squeals, growls, and a relatively small amount of "marginal babbling" consisting of vocal tract closures in sequence with better-formed vowel-like utterances. These syllable-like utterances still differ significantly from adult syllables, e.g. in their durational aspects. At about seven months of age, infants enter the *canonical stage* (also called the *reduplicated babbling stage*; Stark, 1980) where for the first time syllables with adult-like timing characteristics are seen. Many of the infant's utterances during this stage are reduplicated syllables such as "dadada". At the age of approximately 10 months, infants enter a stage known as *variegated* or *nonreduplicated babbling*, characterized by the use of different consonants and vowels within the same babbling sequence (e.g., "badadi"). MacNeilage

and Davis (1990) have hypothesized that the variegated babbling stage is the stage during which infants first begin learning to produce the various phonemes of their native language.

One conclusion that can be drawn from infant babbling data is that many non-speech vocalizations and articulator movements occur well before the onset of frequent speech sounds (e.g., Kaplan and Kaplan, 1971; Oller, 1980; Sachs, 1976; Stark, 1980). In accordance with this view, the simplified babbling process in DIVA occurs in two stages: an early stage during which the sensory-motor relationships of the orosensory-to-articulatory mapping are learned in the absence of speech sounds, and a later stage during which the orosensory targets for each speech sound, encoded by the weights of the phonetic-to-orosensory mapping, are learned. Although relatively rare, speech sounds do occur in the first few months of life; simulations reported in Guenther (1994) verify that including such occurrences during the first stage of babbling in the model does not have a significant adverse effect on the orosensory-to-articulatory learning that takes place during this stage.

Babbling in the model is produced by inducing random movements of the speech articulators. These movements are generated by randomly activating the Articulator Velocity Vector (AVV) cells shown in Figure 2. It should be noted, however, that random movements of the articulators were chosen here for simplicity and generality rather than as an attempt to describe the babbling of infants. Babbling in infants is to a large degree non-random; instead, it appears to be constrained by factors such as neuromotor development and the influence of characteristics of the child's native language (e.g., de Boysson-Bardies, Sagart, and Durand, 1984; de Boysson-Bardies, Halle, Sagart, and Durand, 1989). These constraints presumably make the process of speech sound production learning easier by providing the infant with "training sequences" that are relatively closely related to the movements required in the adult language. For example, the random movements of the model lead to significant sampling of regions of articulator space and orosensory space that are not valid for human languages. Constraints on infant babbling likely aid in limiting infant articulations to more useful portions of the articulator and orosensory spaces. In short, the present work makes no attempt to explain the processes that *generate* babbling, but instead attempts to provide the beginnings of an explanation of how this babbling leads to the tuning of important parameters in the neural mechanisms of speech production by providing a data set which the infant can use to tune these parameters.

The learning processes involved in the two DIVA babbling stages are detailed in Section 3.1 and Section 3.2. With the model simulation operating approximately in real time (as evidenced by the speed of articulator movements visible in a computer animation), the entire babbling sequence takes approximately one hour.

2.2. The Performance Phase

After babbling, the model can produce arbitrary phoneme strings using a set of 29 English phonemes in any combination (see Table 1 for a list of these phonemes). Geometric limitations in the model's simplified articulator system currently prevent learning of a more complete set of English phonemes. In a typical performance, the user will specify a phoneme string for the model to articulate. Performance of the phoneme string can be visualized as follows. The Speech Sound Map (SSM) cell corresponding to the first phoneme in the string is activated. This cell's activity propagates through the phonetic-to-orosensory weights learned during babbling, effectively "reading out" the phoneme's learned orosensory target[2]. The Orosensory Direction Vector (ODV) represents the difference between this target and the current state of the vocal tract; in other words, the ODV codes the desired movement direction in orosensory space. This is then mapped into an appropriate set of articulator velocities. This coordinate transformation is carried out by propagating the ODV activities through the learned weights in the orosensory-to-articulatory mapping. As the articulators move, the shape of the vocal tract, registered through orosensory feedback at the ODV stage, gets closer and closer to the orosensory target for the speech sound. This causes the ODV activity to get smaller and smaller, leading to a slowing and stopping of articulator movements as the target is reached. When ODV activity is sufficiently close to zero (i.e., when the sound has been completed), the SSM cell corresponding to the next phoneme in the string is activated, and the process repeats. These processes are carried out automatically in the neural network defined by the equations in the following paragraphs. The result is a time course of articulator positions that can be viewed as a real-time animation sequence on the computer monitor.

It is important to note that all performance simulations use the same parameter values, learned during a single babbling phase. Furthermore, although no perturbations or con-

2. This statement is simplified for reasons of clarity at this point in the model description. As described in Section 9, the orosensory target depends not only on the current phoneme but also on the targets for later phonemes in the string. This is how anticipatory coarticulation arises.

TABLE 1. Phonemes Learned by the Present Implementation of the DIVA Model

Phoneme	Example	Phoneme	Example	Phoneme	Example
/p/	pin	/š/	ship	/ʌ/	luck
/b/	ball	/ž/	measure	/ɑ/	hot
/t/	tree	/m/	mom	/ɔ/	all
/d/	dog	/n/	nice	/e/	hate
/k/	kick	/ŋ/	sing	/i/	eve
/g/	goal	/l/	lazy	/o/	obey
/θ/	thin	/r/	red	/u/	boot
/ð/	then	/I/	bit	/U/	foot
/s/	sit	/ɛ/	get	/ɝ/	bird
/z/	zebra	/æ/	ash		

NOTE: The simplified articulatory structure of the model allows only a crude mapping between these phonemes and their vocal tract instantiations as learned by the model.

straints to the articulators are encountered during learning, the model exhibits the ability to deal with such constraints automatically during performance, without any new learning (see Section 4). The model also does not train on specific phoneme sequences (cf. the model of Jordan, 1986), but instead learns a context-independent target for each speech sound. The complex context-dependent properties of the articulator movements seen during performance (e.g., contextual variability, carryover coarticulation, and anticipatory coarticulation) arise not from learning what movements to make within these specific contexts (cf. Wickelgren, 1969), but instead are automatic consequences of the shapes of targets learned for the speech sounds and the dynamics of the neural network when producing a string of these sounds. It should also be noted that real speakers typically impose some constraints on which of the possible combinations of phonemes they will use. For example, syllable strings such as /srikp/ feel awkward to produce and are rarely used. No such constraints are implemented in the model, but this is done for the sake of simplicity, not as a prediction about human speech. The model also currently offers no explanations for why such constraints arise in human speech.

2.3. Model Components

The components of the DIVA model are described in the following paragraphs. For clarity of exposition, this discussion will start at the Speech Recognition System block and move clockwise around Figure 2.

Speech Recognition System

During babbling, the Speech Recognition System in the DIVA model interprets the infant's speech signal, activating appropriate cells[3] in the Speech Sound Map whenever the infant produces a speech sound from his/her native language. This can be thought of as an acoustic-to-phonetic mapping. Speech sounds in the present implementation are simply equated to phonemes; the main concepts of the model remain valid, however, for different choices of sound units such as auditory distinctive features. Furthermore, the process of speech recognition is very complex and beyond the scope of this model. Thus, even though the Speech Recognition System is conceptualized as interpreting acoustic signals, no acoustic signal is used in the present implementation. Instead, the Speech Recognition System is implemented as an expert system that looks at key constrictions of the vocal tract to determine which, if any, speech sounds would be produced. If the system recognizes a configuration corresponding to a known speech sound, it activates the corresponding cell in the Speech Sound Map. This activation drives learning in the phonetic-to-orosensory mapping. This corresponds to a situation wherein an infant learns when a match occurs between acoustic effects of his/her own productions and sound categories established by listening to the productions of others.

The process of learning an orosensory target for each speech sound in the present implementation of the DIVA model is currently based on the following assumption: before a normal infant learns to properly and reliably *produce* a given sound, the infant is able to properly and reliably *perceive* that sound. To simplify the simulations, the model starts out with the ability to perceive all of the sounds that it will eventually learn to produce. However, this does not constitute a claim that infants can perceive *all* speech sounds before

3. Each cell, or neuron, in the model corresponds only loosely to an hypothesized population of neurons in the nervous system; the model should thus be considered as a set of hypothesized stages of neural computation rather than as an attempt to identify specific neurons in the brain.

learning to produce *any* speech sounds. It is likely that infants learn to produce some sounds well before they can reliably perceive other sounds. Because learning of the orosensory target for each sound in the model occurs totally independently of the ability to perceive or produce any other sound, the model can similarly learn to produce some sounds before being able to perceive others. It is therefore expected that although the time frames during which infants acquire the abilities to perceive and produce speech sounds overlap substantially, for *any given sound* the ability to reliably perceive the sound develops before the ability to reliably produce it in a normal infant.

Because the current model does not address the self-organization of speech perception, the treatment of the relationship between the development of perception and the development of production is necessarily simplistic: proper perception is simply assumed to have occurred before learning of the production targets begins. The relationship between the development of perception and production skills in infants, however, is at present much less clear. The ability to identify the same phoneme in different contexts and across speakers has been demonstrated at six months of age (Kuhl, 1979), and language-specificity in this phonetic perception has also been demonstrated in six-month-old infants (Kuhl, Williams, Lacerda, Stevens, and Lindblom, 1992). If the learning of phonetic segments begins during the variegated babbling stage as suggested by MacNeilage and Davis (1990), then it would appear that the development of phonetic perception at least begins before the learning of orosensory targets for production. However, infants do produce *some* vowel-like sounds by six months (Oller, 1980; Stark, 1980), and these productions could conceivably play a role in building up the perceptual categories. On the other hand, there is evidence that children with severly limited speech motor abilities can develop relatively normal speech perception (e.g., MacNeilage, Rootes, and Chase, 1967; Rootes and MacNeilage, 1967), while deaf infants typically show large deficits in production without special therapy (e.g., Lynch and Oller, 1989; Oller and Eilers, 1988). Together, these data suggest an important role for proper perception in learning to produce sounds and against an important role for production skills in the development of speech perception. However, they do not clarify whether perceptual *phonetic categories* are in place before the learning of the corresponding production targets begins.

Relatedly, Flege and Eefting (1988) and Flege (1991, 1993) have argued that learners of a second language must establish appropriate phonetic categories before they can reliably

produce the correct phonemes in the second language. The present model's assumption that the perceptual category for a sound exists before the orosensory target for that sound is learned is consistent with this hypothesis. However, some studies suggest a more complicated relationship between perception and production in second language learners. For example, although grouped data in the experiments of Flege (1993) were in accord with the hypothesis that proper perception precedes proper production in second language learners, the data for individual subjects did not support this hypothesis: as many subjects showed large production effects of the second language in the absence of large perception effects as showed large perception effects without production effects.

These data suggest a scenario in which perception and production of a given phonetic segment co-evolve. This view receives support for first language learning from the study of Zlatin and Koenigsknecht (1976), who studied the perception and production of voice onset time (VOT) in two-year-old, six-year-old, and adult subjects. These authors concluded that both perception and production skills continue to improve between ages two and six, with the perceptual status of VOT leading that of production. In terms of the current model, this suggests a learning scenario wherein the Speech Recognition System slowly refines what it considers to be correctly produced examples of each phoneme, and learning of the orosensory targets for production continually "tracks" these changes. Although the present version of the model assumes that perception is reliable and does not change as a consequence of production, future versions of the model that incorporate self-organization in the Speech Recognition System will attempt to more thoroughly address this important issue.

Speech Sound Map

Each cell in this map codes a different speech sound. During babbling, cells in the map are inactive except when the Speech Recognition System determines that the model has produced a speech sound; when this happens, the activity of the corresponding cell in the Speech Sound Map is set to 1. During performance, a higher-level brain center is assumed to sequentially activate the speech sound cells for the desired phoneme string. Thus, the Speech Sound Map cell activities can be summarized as follows:

SSM Activities, Babbling Phase:

$$s_i = \begin{cases} 1 \text{ if recognition system hears i-th sound} \\ 0 \text{ otherwise} \end{cases} \quad (1)$$

SSM Activities, Performance Phase:

$$s_i = \begin{cases} 1 \text{ if production of i-th sound is desired} \\ 0 \text{ otherwise} \end{cases} \quad (2)$$

where s_i is the activity of the cell corresponding to the i^{th} sound, and the index i takes on a value between 1 and 29, corresponding to the 29 phonemes learned by the model.

Orosensory Direction Vector

Orosensory information is key to the DIVA model both for specifying the targets of speech and for activating appropriate articulator movements to reach these targets. Several investigations have hypothesized speech targets within an orosensory frame. Based on the results of bite block experiments showing automatic compensation even on the first glottal pulse (i.e., in the absence of acoustic feedback), Lindblom et al. (1979) hypothesized that "the target of a vowel segment is coded neurophysiologically in terms of its area function by means of corresponding sensory information" (p. 157), where "sensory" here refers to an orosensory reference frame as described above. Similarly, Perkell (1980) posited that acoustic goals are transformed into corresponding orosensory goals during the production process. The task-dynamic model described in Saltzman and Munhall (1989) hypothesizes a vocal tract variable coordinate frame existing between the levels of acoustic goals and motor realization. Because these tract variables characterize key constrictions in the vocal tract, they can be interpreted as another example of sound targets in an orosensory reference frame.

The activities of the Orosensory Direction Vector cells are governed by the following equations:

ODV Activities, Babbling and Performance Phases:

$$d_{j+} = \sum_i s_i z_{ij+} - f_{j+} \tag{3}$$

$$d_{j-} = \sum_i s_i z_{ij-} - f_{j-} \tag{4}$$

where d_{j+} and d_{j-} are the antagonistically paired ODV cell activities corresponding to the j^{th} orosensory dimension, f_{j+} and f_{j-} are antagonistically paired orosensory feedback signals coding position along the j^{th} dimension of orosensory space, s_i is the activity of the i^{th} Speech Sound Map cell, z_{ij+} is the synaptic weight of the pathway from the i^{th} Speech Sound Map cell to the $j+^{th}$ ODV cell, and z_{ij-} is the synaptic weight of the pathway from the i^{th} Speech Sound Map cell to the $j-^{th}$ ODV cell. The weights z_{ij+} and z_{ij-} constitute the phonetic-to-orosensory mapping.

These equations show that ODV cells receive inhibitory tactile and proprioceptive feedback about the state of the vocal tract, represented by the values f_{j+} and f_{j-}. The present implementation uses 11 different orosensory dimensions[4], corresponding to proprioceptive information from individual articulators, tactile information from pressure receptors, and higher-level combinations of information such as the sizes of important constrictions in the vocal tract. A complete list of the orosensory dimensions used in the model is given in Table 2. One of the main tasks of the model during babbling is to differentiate between important and unimportant orosensory cues for a sound. As discussed in Section 3.2, the model successfully extracts the important information for each speech sound from this general set of available sensory information.

Orosensory Direction Vector cells also receive excitatory input via the learned phonetic-to-orosensory mapping; this can be seen as the $\sum s_i z_{ij+}$ and $\sum s_i z_{ij-}$ terms in Equations 3 and 4. When a cell in the Speech Sound Map is activated for performance of the corresponding sound, this input to the ODV acts as a target in orosensory space for producing that sound. The ODV then represents the difference between the learned orosensory target

4. Five orosensory dimensions included in Guenther (1994) have been removed in the current implementation to simplify the simulations. These dimensions, corresponding to individual articulator positions, had no direct bearing on the acoustic properties of the vocal tract and are subsumed in higher-level orosensory dimensions in Table 2.

for the desired sound and the current configuration; this value specifies a desired movement direction in orosensory space that is then mapped into a set of articulator velocities to move the vocal tract in this direction.

TABLE 2. Orosensory Dimensions in the Present Implementation of the DIVA Model.

Tongue body horizontal position with respect to maxilla
Tongue body height with respect to maxilla
Tongue body pressure receptors
Tongue tip horizontal position with respect to maxilla
Tongue tip height with respect to maxilla
Tongue tip pressure receptors
Lip protrusion
Lip aperture
Lower lip pressure receptors
Upper lip pressure receptors
Velum height

NOTE: Most of these dimensions are closely related to the tract variables of Saltzman and Munhall (1989).

During the first stage of babbling, changes in the configuration of the vocal tract will cause changes in the Orosensory Direction Vector activities. These changes drive learning in the orosensory-to-articulatory mapping as described in Section 3.1. Note that since no speech sounds are produced during the first babbling stage, all s_i are zero and no excitatory input propagates to the ODV cells. During the second babbling stage, random production of a speech sound will result in activation of the corresponding s_i. Now, ODV cell activity reflects the difference between the current vocal tract configuration (from the f_{j+} and f_{j-}) and the orosensory target for that speech sound (encoded by the weights z_{ij+} and z_{ij-}). This difference drives learning in the phonetic-to-orosensory mapping as described in Section 3.2.

Articulator Velocity Vector

The Articulator Velocity Vector consists of a set of cells that command movements of the articulators. The activity of each cell is meant to correspond roughly to a commanded contraction of a single muscle or a group of muscles in a fixed synergy. The cells are formed into antagonistic pairs, with each pair corresponding to a different degree of freedom of the articulatory mechanism. Table 3 tabulates the articulatory degrees of freedom used in the model.

TABLE 3. Articulatory Degrees of Freedom in the Present DIVA Implementation

Raise/lower jaw
Raise/lower tongue body with respect to jaw
Raise/lower tongue tip with respect to tongue body
Raise/lower upper lip
Raise/lower lower lip with respect to jaw
Raise/lower velum
Forward/backward extension of tongue body with respect to jaw
Forward/backward extension of tongue tip with respect to tongue body
Forward/backward extension of both lips simultaneously

During babbling, Articulator Velocity Vector cells are randomly activated to produce movements of the articulators. It is assumed that this occurs via an endogenous random generator that overrides other AVV inputs during babbling (see Bullock et al., 1993; Gaudiano and Grossberg, 1991). During performance, activation of the Articulator Velocity Vector cells occurs through the phonetic-to-orosensory and orosensory-to-articulatory mappings. Specifically, AVV cell activities are governed by the following equations:

AVV Activities, Babbling Phase:

$$a_{k+} = \begin{cases} 1 \text{ with probability } 1/3 \text{ for each trial} \\ 0 \text{ otherwise} \end{cases} \quad (5)$$

$$a_{k-} = \begin{cases} 1 \text{ with probability } 1/3 \text{ for each trial} \\ 0 \text{ otherwise} \end{cases} \quad (6)$$

AVV Activities, Performance Phase:

$$a_{k+} = \sum_j [d_{j+}]^+ w_{j+k+} + \sum_j [d_{j-}]^+ w_{j-k+} \quad (7)$$

$$a_{k-} = \sum_j [d_{j+}]^+ w_{j+k-} + \sum_j [d_{j-}]^+ w_{j-k-} \quad (8)$$

where a_{k+} and a_{k-} are the antagonistic pair of activities corresponding to the k^{th} articulatory degree of freedom, w_{j+k+} is the synaptic weight projecting from the $j+^{th}$ ODV cell to the $k+^{th}$ AVV cell (with analogous definitions for the various +, - combinations), and $[x]^+$ is a rectification function such that $[x]^+ = 0$ for $x < 0$ and $[x]^+ = x$ for $x \geq 0$. The weights w_{j+k+}, w_{j+k-}, w_{j-k+}, and w_{j-k-} make up the orosensory-to-articulatory mapping.

The transformation performed by the orosensory-to-articulatory mapping can be envisioned as a transformation of the movement specification from a sensory coordinate frame to a motor coordinate frame. As described above, the ODV cells form a vector in orosensory coordinates coding the distance and direction from the current vocal tract configuration to the target region. Multiplying this vector by the matrix of weights in the orosensory-to-articulatory pathways (Equations 7 and 8) effectively produces a vector describing the distance and direction of desired movement in the motor coordinate frame. This vector serves as the basis for commanded velocities of the articulators as described in the next paragraph.

GO Signal

The GO signal (Bullock and Grossberg, 1988) is used to multiplicatively gate the movement commands at the Articulator Velocity Vector before sending them to the motoneurons controlling the contractile state of the muscles. This signal corresponds to volitional control of movement onset and speed in a human being and is discussed within the context of speaking rate in Section 7. The equation governing articulator velocities is as follows:

Articulator Velocities:

$$v_k = G \times [a_{k+} - a_{k-}] \tag{9}$$

where v_k is the velocity along the k^{th} articulatory degree of freedom and G is the value of the volitional GO signal (varying between 0 for minimum speaking rate and 1 for maximum speaking rate). The GO signal is fixed at a value of 0.5 during babbling.

3. Acquisition of Speaking Skills

Acquisition of speaking skills in DIVA consists of finding appropriate parameters, or synaptic weights, for the phonetic-to-orosensory and orosensory-to-articulatory mappings during the two stages of the babbling phase. The learning processes involved during babbling are described in the following paragraphs.

3.1. Developing Coordinative Structures: The Orosensory-to-Articulatory Mapping

In the first stage of babbling, the DIVA model learns a mapping from directions in orosensory space (coded by the ODV stage) to movement directions in articulator space (coded by the AVV stage). A portion of this mapping is shown in Figure 3. Learning of the orosensory-to-articulatory mapping occurs as follows. Randomly activated Articulator Velocity Vector cells cause movements of the speech articulators which are reflected through orosensory feedback as changes in activity of the Orosensory Direction Vector cells. It is these *changes* in ODV activities, rather than the magnitude of activities, that drive learning in the orosensory-to-articulatory pathways according to the following equations:

$$\frac{d}{dt} w_{j+k+} = \varepsilon_1 a_{k+} \left(- \alpha_1 w_{j+k+} - \frac{d}{dt} d_{j+} \right) \tag{10}$$

$$\frac{d}{dt} w_{j+k-} = \varepsilon_1 a_{k-} \left(- \alpha_1 w_{j+k-} - \frac{d}{dt} d_{j+} \right) \tag{11}$$

$$\frac{d}{dt} w_{j-k+} = \varepsilon_1 a_{k+} \left(- \alpha_1 w_{j-k+} - \frac{d}{dt} d_{j-} \right) \tag{12}$$

$$\frac{d}{dt}w_{j\text{-}k\text{-}} = \varepsilon_1 a_{k\text{-}}\left(-\alpha_1 w_{j\text{-}k\text{-}} - \frac{d}{dt}d_{j\text{-}}\right) \tag{13}$$

where ε_1 is a learning rate parameter and α_1 is a learning decay parameter. Thus, a decrease in an ODV cell's activity results in an increase in the weight projecting from the ODV cell to active Articulator Velocity Vector cells; these AVV cells are responsible for the movements that resulted in the initial decrease of ODV activity. In this way, each ODV cell learns a set of articulator velocities that cause movements to decrease the ODV cell's activity, i.e. articulator movements that move the vocal tract in the desired direction.

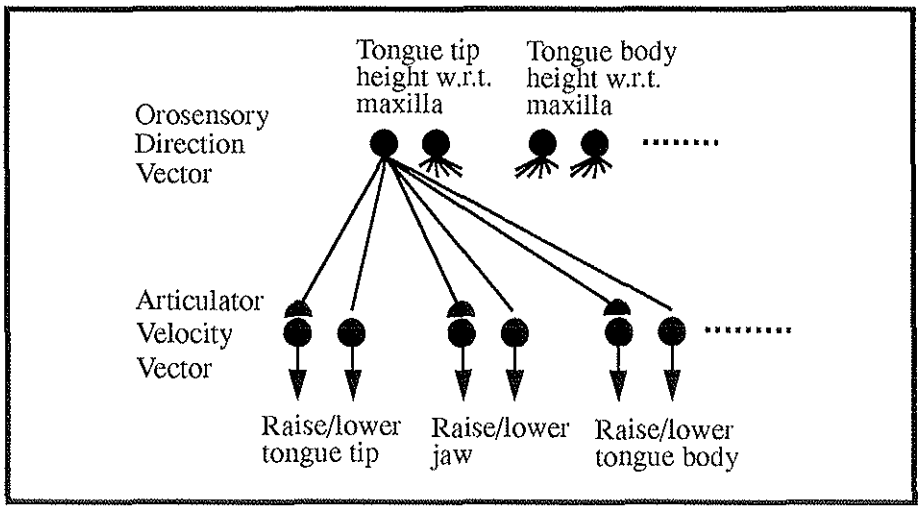

FIGURE 3. Schematized view of a portion of the orosensory-to-articulatory mapping after babbling. ODV cells, each coding a desired movement direction in orosensory space, project with large weights to AVV cells that move the vocal tract in the appropriate direction. Projections to other AVV cells have withered away to zero during learning. Activity at an ODV cell during performance will propagate through the large weighted pathways and activate the corresponding set of articulator movements; this set of articulator movements constitutes a coordinative structure.

The mapping between orosensory variables and articulator variables is analogous to the mapping between vocal tract variables and articulator variables in the task-dynamic model of Saltzman and Munhall (1989), and both are related to the redundant inverse kinematics problem of robotics (e.g., Craig, 1986). Whereas the mapping in DIVA is learned, the mapping in the task-dynamic model is solved mathematically by calculating a weighted Jacobian pseudoinverse and adding terms to provide a neutral attractor (see Sections 8 and 9 for brief discussions of the neutral attractor) and to prevent unwanted movements after an orosensory target has been reached (a common problem of pseudoinverse tech-

niques). The resulting equation relating articulator movements to orosensory variables is very complex; in fact, Munhall, Ostry, and Flanagan (1991; p. 305) state that the complexity of this mapping is one reason for looking to simpler coordinate frames for movement planning, such as joint coordinates. However, the inverse kinematics mapping in DIVA is very simple (characterized by Equations 7 and 8) and the parameters defining the mapping are easily learned[5]. Furthermore, Guenther (1992) and Bullock et al. (1993) show how a direction-to-velocity inverse kinematics approach like the one used in DIVA leads to motor equivalence properties that are very difficult to explain with a joint coordinate planning approach.

The orosensory-to-articulatory mapping in DIVA is also closely related to the *coordinative structure* modeling concept (e.g., Easton, 1972; Fowler, 1980; Kelso et al., 1984; Saltzman and Kelso, 1987). A coordinative structure is a task-specific grouping of articulators. For example, Kelso et al. (1984) report that when a subject's jaw is perturbed during the production of /b/, compensation is seen in the movements of the upper and lower lips but not movements of the tongue. When perturbation is applied during /z/ production, however, compensation is seen in the movements of the tongue but not movements of the lips. Thus, it appears that these subjects use a coordinative structure consisting of the upper lip, lower lip, and jaw when the task is to produce a /b/, and a coordinative structure consisting of the tongue and jaw when the task is to produce a /z/. Such groupings arise naturally in the DIVA self-organization process. Figure 3 schematizes the results after babbling for the ODV cell coding an increase in tongue tip position with respect to the maxilla. This cell now projects through large weights to AVV cells that raise the tongue tip, the jaw, and the tongue body; the weights for projections to other AVV cells have withered to zero. During performance, a positive activity at this ODV cell will arise when the "task" is to increase tongue tip constriction degree, as for a /z/. This positive activity will propagate through the pathways with large weights (see Equations 7 and 8), resulting in the simultaneous raising of the tongue tip, tongue body, and jaw; this task-specific grouping of articulator movements constitutes a coordinative structure. If one of these three movements is

5. Because the inverse kinematic mapping in DIVA is the result of a learning process rather than an explicit calculation, it is not possible to precisely characterize this mapping, e.g. in terms of a Jacobian pseudoinverse. Instead, the mapping can best be characterized as an approximate Jacobian pseudoinverse whose exact form is the result of complex dynamic interactions involving the training sequence and the learning laws of Equations 10-13.

blocked (e.g., a bite block could be used to prevent jaw movement), the other movements continue to decrease tongue tip constriction degree, resulting in the automatic compensation demonstrated in the model simulations of Section 4. As the tasks change to produce different phonemes, different ODV cells will have positive activity, leading to different coordinative structures for producing the required movements. In this way, the model automatically marshals only appropriate coordinative structures, as seen in the human speaking data of Kelso et al. (1984).

3.2. Learning Sound Targets: The Phonetic-to-Orosensory Mapping

The synaptic weights in the pathways projecting from a Speech Sound Map cell to the Orosensory Direction Vector cells represent a vocal tract target for the corresponding speech sound in orosensory space. When the changing vocal tract configuration is identified by the Speech Recognition system as producing a speech sound during the second stage of babbling, the appropriate Speech Sound Map cell's activity is set to 1. This gates on learning in the synaptic weights of the phonetic-to-orosensory pathways projecting from that cell, and, as described in the following paragraphs, this allows the model to modify the orosensory target for the speech sound based on the current configuration of the vocal tract as seen through orosensory feedback at the ODV stage.

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

The manner in which targets are learned in DIVA provides a unified answer to these questions. Figure 4 schematizes the learning sequence for the vowel /i/ along two dimensions (corresponding to lip aperture and tongue body height) of orosensory space. The first time

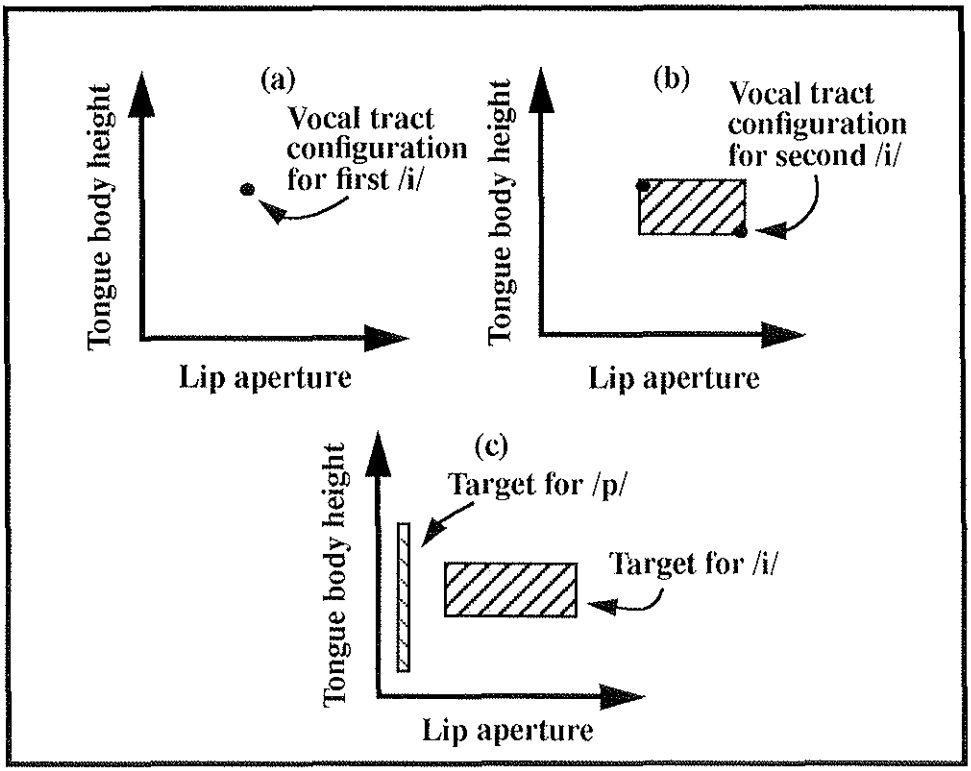

FIGURE 4. Learning of the convex region target for the vowel /i/ along orosensory dimensions corresponding to lip aperture and tongue body height. (a) The first time /i/ is produced during babbling, the learned target is simply the configuration of the vocal tract when the sound was produced. (b) The second time /i/ is babbled, the convex region target is expanded to encompass both vocal tract configurations used to produce the sound. (c) Schematized convex regions for /i/ and /p/ after many productions of each sound during babbling. Whereas the target for /i/ allows large variation along the dimension of lip aperture, the target for the bilabial stop /p/ requires strict control of this dimension, indicating that the model has learned that lip aperture is an important aspect of /p/ but not /i/.

the phoneme is produced during babbling, the corresponding cell in the Speech Sound Map learns the orosensory position that caused the phoneme. This corresponds to a point in orosensory position space, schematized in Figure 4a. The next time the phoneme is babbled, the Speech Sound Map cell expands its learned target to be a convex region that encompasses both the previous orosensory position and the current orosensory position, as shown in Figure 4b; this occurs via the simple and biologically plausible learning law of Equations 14 and Equations 15 below. In this way, the model is constantly expanding its convex region target for /i/ to encompass all of the various vocal tract configurations that can be used to produce /i/.

Now we can address the questions posed above. Consider the convex regions that result after many instances of producing the vowel /i/ and the bilabial stop /p/ (Figure 4c). The convex region for /p/ does not vary over the dimension of lip aperture but varies largely over the dimension of tongue body height; this is because all bilabial stops that the model has produced have the same lip aperture, but tongue body height has varied. In other words, the model has learned that bilabial aperture is the important orosensory invariant for producing the bilabial stop /p/. Furthermore, whereas lip aperture is the important orosensory dimension for /p/, the model has learned that this dimension is not very important for /i/, as indicated by the wide range of lip aperture in the target for /i/ in Figure 4c. Finally, since convex region learning relies on language-specific recognition of phonemes by the infant, the shapes of the resulting convex regions will vary from language to language.

The neural mechanism used to learn the convex region targets in DIVA is related to the Vector Associative Map detailed in Gaudiano and Grossberg (1991). The learning laws governing modification of the synaptic weights are:

$$\frac{d}{dt}z_{ij+} = \varepsilon_2 s_i\left(\alpha_2 z_{ij+} - [d_{j+}]^+\right) \tag{14}$$

$$\frac{d}{dt}z_{ij-} = \varepsilon_2 s_i\left(\alpha_2 z_{ij-} - [d_{j-}]^+\right) \tag{15}$$

where ε_2 is a learning rate parameter, α_2 is a learning decay parameter, and $[x]^+$ is a rectification function as defined earlier. The learning laws of Equations 14 and 15 ensure that modification of a given phoneme's orosensory target only occurs when that phoneme is being produced. The weights start out large (initialized to 1.0) and primarily decrease with learning; this decrease in the weights corresponds to an increase in the size of the orosensory convex region target.

To see why this is the case, refer to Figure 5, which schematizes the mapping from a Speech Sound Map cell to the antagonistic pair coding one dimension of the Orosensory Direction Vector. The orosensory feedback signal antagonistic pairs (f_{j+}, f_{j-}) each sum to a constant value of 1; this kind of push-pull relationship between cell activities is often found in the nervous system (e.g., Sakata et al., 1980). Assume a large value of ε_2 and a very small value of α_2 in Equations 14 and 15. The first time the speech sound corresponding to s_i is produced during babbling, the weight pair (z_{ij+}, z_{ij-}) will converge to

the value of (f_{j+}, f_{j-}) when this sound occurred. Assume that this occurred with $(f_{j+}, f_{j-}) = (0.4, 0.6)$. From Equations 7 and 8 it is clear that during performance only positive d_{j+} and d_{j-} will activate articulator movements. With $(z_{ji+}, z_{ji-}) = (0.4, 0.6)$, from Equations 3 and 4 we can see that any value of (f_{j+}, f_{j-}) other than (0.4, 0.6) will drive an articulator movement when s_i is activated to 1. This corresponds to a point attractor or point target at (0.4, 0.6) for (f_{j+}, f_{j-}).

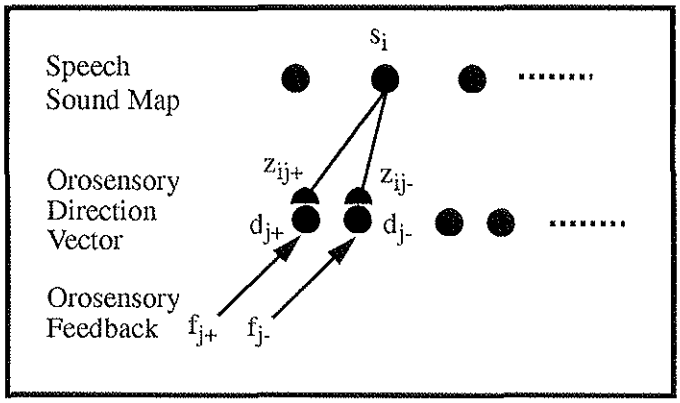

FIGURE 5. Portion of the phonetic-to-orosensory mapping from a Speech Sound Map cell to the antagonistic pair coding one dimension of orosensory space.

Now consider what happens if the sound corresponding to s_i is produced a second time, with $(f_{j+}, f_{j-}) = (0.5, 0.5)$. Learning will drive the weights (z_{ij+}, z_{ij-}) to (0.4, 0.5). With this weight pair, we see from Equations 3 and 4 that a positive d_{j+} or d_{j-} will only result if (f_{j+}, f_{j-}) is outside the range $(0.4 \le f_{j+} \le 0.5, 0.5 \le f_{j-} \le 0.6)$. This range thus defines a convex region attractor. Further decreases in the weight values will result in further increases in the size of the convex region attractor.

An interesting property of this learning process is that the model can learn to "ignore" totally unimportant orosensory dimensions by allowing variability throughout the entire range of such dimensions. This reduces the need for the nervous system to include only the most important orosensory dimensions in the speech sound target specifications. For example, little harm is done by including orosensory dimensions that are important only for some languages but not for others, since speakers of languages that do not use a dimension can simply learn to ignore it. Despite this added flexibility, it is quite possible that the neural transformation from vocal tract tactile and proprioceptive information into the orosensory dimensions used for target specification is an adaptive one that "chooses"

the most important dimensions for a particular language. This adaptability is not included in the current version of the model, and future research will explore the use of self-organizing mappings to perform this transformation.

The convex region theory constitutes a new entry in the long-standing debate in the speech production literature over the nature of the "targets" as specified to the production mechanism (see Levelt, 1989, chapter 11 for a recent review). Early researchers proposed spatial targets for the articulators (Henke, 1966) and muscle length targets (e.g., Cohen, Grossberg, and Stork, 1988; MacNeilage, 1970); unfortunately, these models cannot account for compensatory movements of one articulator when another articulator cannot reach its "normal" position (e.g., Abbs and Gracco, 1984; Folkins and Abbs, 1975; Kelso et al., 1984; Lindblom et al., 1979). To overcome this, later models hypothesized that the targets are more abstract functions of the vocal tract shape that correspond more closely to the speech signal (e.g., Lindblom et al., 1979; Perkell, 1980; Saltzman and Munhall, 1989). A common assumption of these models is that targets correspond to (possibly context-dependent or time-varying) canonical *positions* of articulators or vocal tract variables. In contrast, Keating (1990) hypothesized a "window theory" of coarticulation wherein the target for each articulator is not a fixed position, but a range of possible positions. As Fowler (1990) points out, however, in many cases the position of a single articulator may vary because this articulator is used in concert with other articulators to produce a higher-level goal which does *not* show much variability. For example, Abbs and Netsell (1973; see also Abbs, 1986) report that whereas large variability is seen in lower lip height and jaw height during production of the vowel /a/, the quantity [lower lip height + jaw height] remains relatively constant. Variability is also seen in lower lip and upper lip heights used to produce bilabial closure (e.g., Kelso et al., 1984). In this case, it is insufficient to simply move the articulators to the acceptable ranges for upper lip height and lower lip height; in addition, one must insure that the resulting lip aperture is zero. A simple window theory as proposed by Keating (1990) cannot explain these data.

The current theory handles these shortcomings. Within this theory, the target for a speech sound is specified in a high-dimensional orosensory space. This orosensory space includes tactile information from pressure receptors and more complex information corresponding to higher-order combinations of tactile and proprioceptive information, such as the degree of constriction at different points along the vocal tract (see Table 2). Each dimension of

the orosensory target specifies a range of acceptable positions along that dimension. The babbling process causes the system to learn very small target ranges for acoustically important orosensory dimensions and large ranges for unimportant dimensions, thus insuring proper production despite allowing large variability in unimportant dimensions.

The preceding paragraphs have described the process by which the DIVA model learns to produce speech sounds. The remainder of this article investigates the properties of the articulator movements during performance of phoneme strings. These properties arise largely as a result of the nature of speech targets and mappings between coordinate frames learned during the babbling phase.

4. Motor Equivalence

The direction-to-velocity nature of the orosensory-to-articulatory mapping in DIVA provides the model with the ability to automatically compensate for perturbations or constraints on articulator movements despite the fact that the model never encounters such constraints during learning. Guenther (1992) and Bullock et al. (1993) discuss in detail how these motor equivalence properties arise in a direction-to-velocity mapping, but not in other forms of inverse kinematic mappings, for goal-directed reaching using a multi-joint arm, and Guenther (1994) details the motor equivalence properties of DIVA. Simulation results verifying these properties are very briefly summarized in this section for completeness.

Figure 6 shows the configurations reached by the model for /p/ in the phrase "sap" under several different conditions. In Figure 6a, the configuration reached during normal, unperturbed speech is shown. In Figure 6b, a perturbation has been applied to the lower lip during /p/ production. As in human subjects (e.g., Abbs and Gracco, 1984), the upper lip compensates by moving further down to make contact with the lower lip for the bilabial closure. In Figure 6c, a perturbation has been applied to the jaw during /p/ production. Here, the lower and upper lips compensate by moving further to make the bilabial closure, as reported experimentally (e.g., Folkins and Abbs, 1975; Kelso et al., 1984). Finally, Figure 6d shows the result of fixing the jaw open during production of the phrase, as would occur if a bite block were held between the teeth while speaking (e.g, Lindblom,

Lubker, and Gay, 1979). Again, upper and lower lips successfully compensate for the loss of jaw movement.

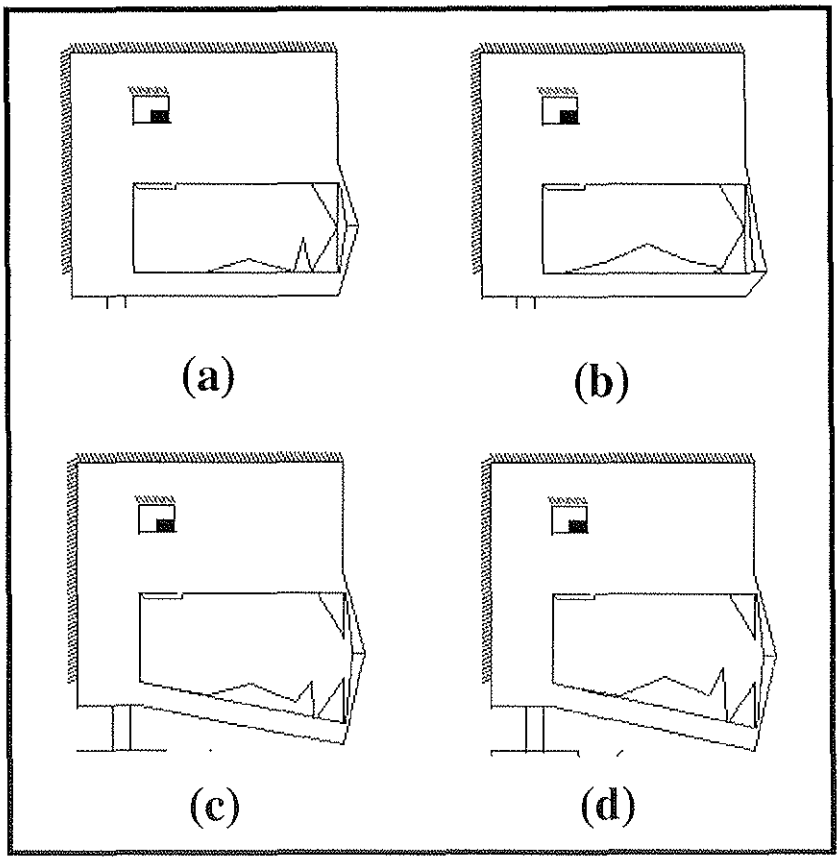

FIGURE 6. Motor equivalence simulation results. Each figure shows a snapshot of the model's articulator configuration during the bilabial closure for /p/ in the word "sap" under a different condition. (a) Normal speech. (b) Downward perturbation to the lower lip during /p/ production. (c) Downward jaw perturbation during /p/ production. (d) Fixed jaw during entire utterance. The model automatically compensates for the constraints in each case despite never having encountered any such constraints during learning.

5. Direct Relationship Between Velocity and Distance

A widely reported characteristic of speech articulator movements is that peak articulator velocity varies directly with magnitude of articulator displacement (e.g., Houde, 1967; Kent and Moll, 1972a,b; Kozhevnikov and Chistovich, 1965; Kuehn and Moll, 1976; Perkell, 1969; Sussman and Smith, 1971). This property has been credited with producing

nearly constant movement durations independent of movement extent (MacNeilage, 1970; Fowler 1980).

Investigation of Equations 7-9 reveals that for a given speaking rate, peak articulator velocity in DIVA will vary directly with ODV activity. Because the ODV activity codes the difference between the current vocal tract configuration and the orosensory target, it can be predicted that peak articulator velocity in DIVA will indeed vary directly with magnitude of articulator displacement. This property is not obvious, however, for several reasons. For example, the distance coded by ODV activity is defined in orosensory coordinates rather than articulator coordinates, and the activity of many ODV cells can simultaneously affect the velocity of a single articulator (e.g. jaw raising can be commanded to different degrees by an ODV cell coding lip aperture and an ODV cell coding tongue body height for production of a single segment). Because of these complicating factors, a simulation was run to determine the relationship between peak velocity of tongue dorsum movement and tongue dorsum displacement over a range of phonemes and contexts. The results of this simulation are shown in Figure 7 (bottom half), along with data from human speakers (top half). The top part of this figure shows data for the tongue dorsum of a speaker in the study of Ostry and Munhall (1985) while producing various vowels in /kVkV/ sequences at both a fast rate and a slow rate. The bottom part of the figure shows corresponding results from DIVA simulations. In agreement with the Ostry and Munhall data and the other experimental studies mentioned above, a direct relationship is seen between peak velocity and articulator displacement in the DIVA simulations. Furthermore, a systematic increase in the slope of this relationship is seen with an increase in speaking rate in both the Ostry and Munhall data and the DIVA simulations. This speaking rate effect is addressed further in Section 7.

It should be noted here that the units for distance and velocity in the model simulations are rather arbitrary, typically relating to pixel sizes, cell activations, or time step sizes. These units are linearly related to "real world" units such as inches and seconds. This is sufficient because only relative magnitudes are of importance for the purposes of this article. Because of their relatively arbitrary nature, the units are not stated in tables and graphs.

A final interesting result concerning the velocity/distance relationship of articulator movements comes from a comparison of fricative and stop consonants. Kuehn and Moll (1976) note that the slope of the velocity/distance relationship was larger for movements toward

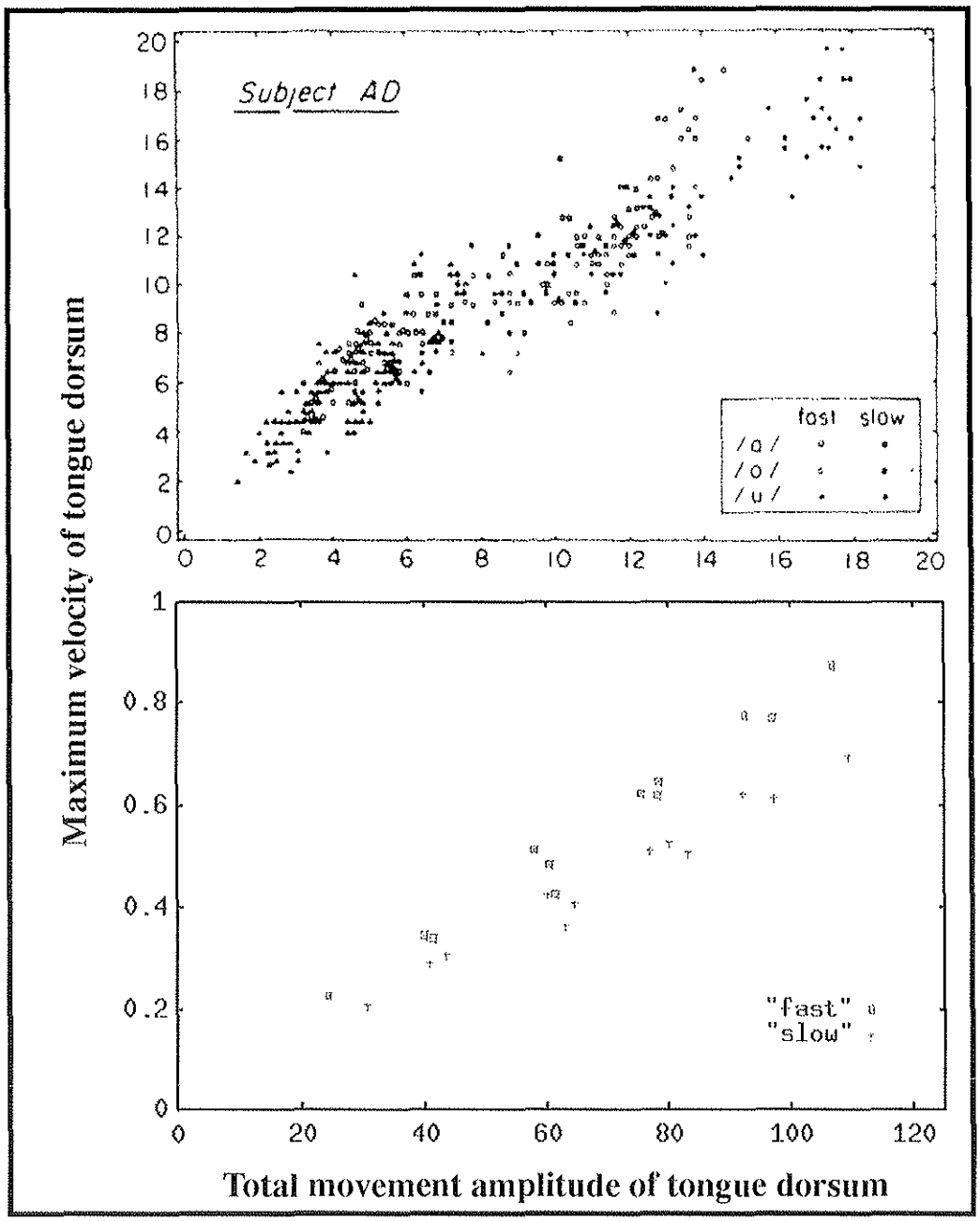

FIGURE 7. Direct relationship between maximum movement velocity and movement amplitude. Top figure shows data for the tongue dorsum of a speaker in the study of Ostry and Munhall (1985) while producing various vowels in /kVkV/ sequences [Reprinted from Ostry and Munhall, 1985]. Bottom figure shows corresponding results in DIVA simulations. In agreement with a large number of experimental studies, a direct relationship is seen between peak velocity and articulator displacement. Furthermore, a systematic change in the slope of this relationship is seen with an increase in speaking rate in both the Ostry and Munhall data and the DIVA simulations.

stops than for movements toward fricatives. That is, for a fixed movement distance of the occluding articulator, movements toward a stop are faster than movements toward a fricative. This result is explained by the DIVA model because of the following property: the orosensory targets for stops include larger target levels of pressure receptor activity than do fricative targets. This will lead to a larger positive ODV activity along the dimension corresponding to the pressure receptor for a stop, and this will subsequently add to the total amount of AVV activity and articulator speed. This property was verified by having the model produce the words "pat" and "path" with the same movement distance required for the tongue tip occlusion in the two cases. For the stop /t/ in "pat", the maximum velocity of the tongue tip was 0.88 (again, the units are arbitrary distance units), and for the fricative /θ/ in "path", the maximum velocity was 0.74. Thus the model not only reproduces the widely reported direct relationship between maximum velocity and distance for articulator movements, but it also accounts for differences in the slope of this relationship for different segment classes.

6. Variability in Place of Articulation

The existence of target ranges along orosensory dimensions in DIVA, rather than explicit target positions, predicts that variability will be seen in the place of articulation along these dimensions. This is because no movements are commanded for positions anywhere within the target range, so entering the range at different positions during different production trials (due, for example, to contextual or biomechanical influences) will lead to different places of articulation. Furthermore, because the size of the target range along an orosensory dimension reflects the amount that the vocal tract is allowed to vary along that dimension while still adequately producing the same phoneme, more variation will occur for acoustically less important dimensions.

An example of this phenomenon in human speech comes from studies of place of articulation for velar stops. English speakers/hearers do not differentiate between velar and palatal stop consonants; as a result, wide anteroposterior variability is seen in the place of constriction for the stop consonants /k/ and /g/ in different vowel contexts (e.g, Daniloff, Schuckers, and Feth, 1980; Kent and Minifie, 1977). Kent and Minifie point out that if the target position for /k/ or /g/ is very concrete and positionally well-defined, then the variation cannot be explained by a target position model. Furthermore, if the target positions

are only loosely defined, the possibility exists for too much variation that can destroy phonemic identity. Since large anteroposterior variation is seen in /k/ and /g/ but little or no variation is allowable in the vertical position of the tongue body (i.e., the tongue body must contact the palate), it appears that neither a well-defined nor loosely defined target position will suffice. Instead, it appears that tongue body target *ranges* are defined separately for anteroposterior position and vertical position, with a large target range for the former and a much smaller range for the latter. This is captured by the shape of the convex region target learned for /k/ by DIVA (see Figure 12a), and simulations of this phenomenon as a result of carryover coarticulation and anticipatory coarticulation are given in Section 8 and Section 9, respectively.

For consonants, it is clear that humans must strictly control the place of articulation along the orosensory dimension corresponding to the constriction degree. For vowels, however, it is unlikely that any orosensory dimension need be so strictly controlled (e.g., Lindblom, 1963). Still, the model predicts that more variability will be seen for vowels along acoustically less important dimensions. The hypothesis of more articulatory variability along acoustically less important dimensions for the vowels /i/ and /a/ was tested on human subjects in studies by Perkell and Nelson (1982, 1985). These reports showed more variability in tongue position along a direction parallel to the vocal tract midline than for the acoustically more important tongue position along a direction perpendicular to the vocal tract midline, supporting this hypothesis. A simulation of this property in DIVA is shown in the bottom half of Figure 8. For this simulation, repeated utterances of /i/ in different contexts and at different rates leads to the scatter of tongue body positions (indicated by small black squares) in the figure. Clearly, variation along the acoustically more important dimension of vertical tongue body position (i.e., position in the direction perpendicular to the midline of the vocal tract) is smaller than variation along the acoustically less important dimension of horizontal position of the tongue body; the corresponding result of a human subject in the study of Perkell and Nelson (1982; 1985) is shown in the top part of Figure 8 ("MID" pellets correspond to tongue body in DIVA). This occurs in the model because the speech recognition system "hears" /i/ when the tongue body occupies a relatively large range of positions along the dimension of tongue body horizontal position but a relatively small range of positions along the dimension of tongue body vertical position, leading the model to learn a convex region target for /i/ with this shape. During production, the actual position on this convex region achieved for /i/ will vary depending on con-

text and rate, leading to a scatter of positions that approximates the shape of the learned target as seen in Figure 8.

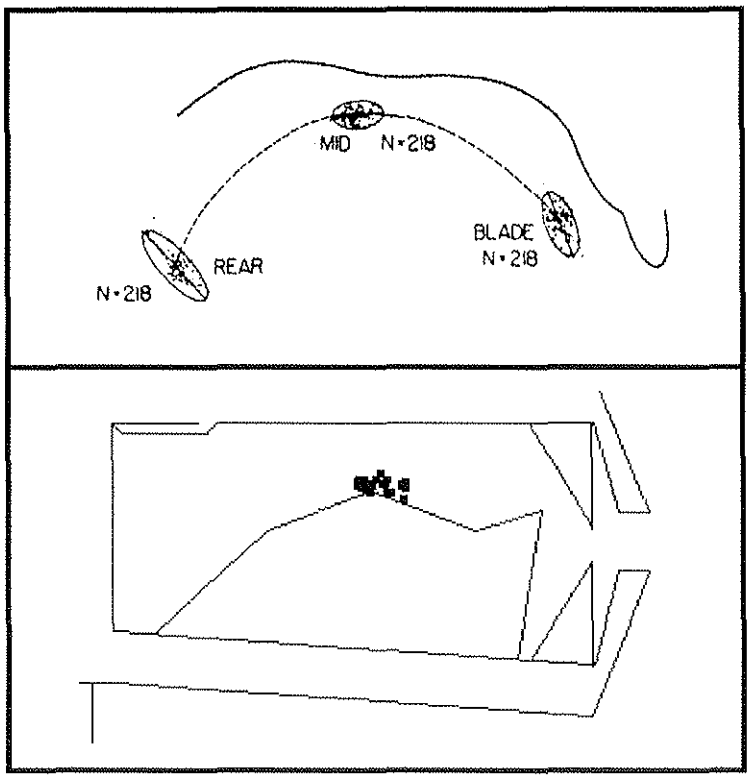

FIGURE 8. Variability in place of articulation for the vowel /i/ in when spoken under various conditions. Top figure is data for a speaker in a study by Perkell and Nelson (adapted from Perkell and Nelson, 1985). Bottom figure shows the corresponding results from a DIVA simulation. In both cases, variation along the acoustically more important dimension of tongue body vertical position is smaller than along the acoustically less important dimension of tongue body horizontal position.

A final example of variability of place of articulation in DIVA comes from the observation that in some cases one should see very wide, but not complete, variation along an orosensory dimension that is largely, but not completely, irrelevant for a particular speech sound. For example, lip aperture is relatively unimportant for velar, alveolar, and dental consonants, but the lips cannot be completely closed during their production. Correspondingly, wide variation of lip aperture can be observed for these sounds, but not complete closure of the lips. Again, such an observation is very difficult to explain using a target position model.

An interesting example of this phenomenon comes from studies of velum position during vowel production. Production of vowels in different consonant contexts results in large, but not complete, variability in velum position during the vowel (Kent, Carney, and Severeid, 1974). For example, if a vowel is produced between two non-nasal consonants as in the word "dad", the velum remains completely closed throughout the utterance. When a vowel is produced between a nasal and a nonnasal consonant as in the word "dan", the velum smoothly transitions from closed to open during the vowel. From these observations it might appear that no fixed target velum position is specified for vowels. However, Kent et al. (1974) report that for a vowel between two nasal consonants, a slight but incomplete raising of the velum occurs during the vowel, followed by a lowering of the velum for the final nasal consonant. As Keating (1990) points out, these data provide a compelling case for a target range from maximally closed to largely, but not completely, open, rather than for any canonical target position.

A DIVA simulation of these data is illustrated in Figure 9. The squares in this figure indicate the velum position while producing the phonemes in the phrase "dad". Here, it is clear that the velum remains closed during the entire utterance. The circles in the figure show the velum positions while producing the phrase "man". Here, we see the velum raising slightly, but not completely, during production of /a/ before lowering again for the final /n/, as reported by Kent et al. This occurs in DIVA because the model has learned a range of acceptable velum positions for the vowel rather than a particular velum position, and the velum is moved to the closest position along that range. In a non-nasal consonant context the closest position in the range is a closed velum position, and in a nasal consonant context the closest position is a largely but not maximally open velum.

7. Speaking Rate Effects

Much research in the past twenty years has investigated how changes in speaking rate affect the production of speech sounds (e.g., Adams, Weismer, and Kent, 1993; De Nil and Abbs, 1991; Flege, 1988a; Gay et al., 1974; Gopal, 1990; Kuehn, 1973; Kuehn and Moll, 1976; Ostry and Munhall, 1985). A common result from these studies is that changes in speaking rate have differential effects for the movements corresponding to vowels and consonants: increasing rate causes an increase in the velocities of movements corresponding to consonantal gestures, but it causes less of an increase, or even a decrease, in the

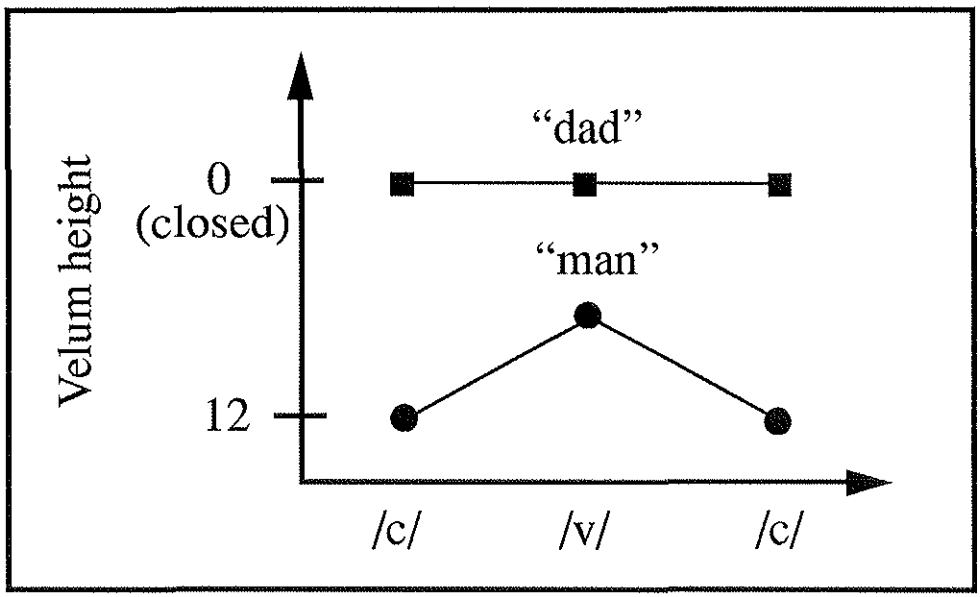

FIGURE 9. Simulation result showing variability of velum position during vowel production in different consonant contexts. The squares show velum position during production of the word "dad"; here, the velum remains completely closed during the vowel. The bottom row shows production of the word "man". In this case, the velum raises slightly but not completely for the vowel, as reported for human subjects (Kent, Carney, and Severeid, 1974).

velocities of movements corresponding to vowel gestures (e.g., MacNeilage and Ladefoged, 1976, pp. 99). This seems to indicate different control strategies for vowels and consonants, and these data have reasonably been used to support a coproduction model of coarticulation positing different underlying control structures for the two sound types (e.g., Fowler, 1980, pp. 121-2, 128).

Several researchers have also noted that different speakers tend to use different strategies to increase speaking rate (Kuehn, 1973; Kuehn and Moll, 1976; Ostry and Munhall, 1985): some speakers rely more on increases in velocity, and others rely more on decreases in movement amplitudes. These will be referred to as the *velocity strategy* and *amplitude strategy*, respectively.

The velocity strategy is already inherent to DIVA as a consequence of the GO signal that gates movement commands. In the VITE model of trajectory formation (Bullock and Grossberg, 1988), the GO signal is a volitional signal embodying the will to move at a particular speed; increased movement speed is achieved by increasing the GO signal, which

in turn multiplicatively gates desired movement direction commands. This is carried out in DIVA by Equation 9. Other things being equal, increasing the GO signal in this equation directly increases articulator velocities and, therefore, speaking rate. The multiplicative relationship between the GO signal and a desired movement vector as exemplified by Equation 9 has been used to explain a very large amount of data from the movement control literature (Bullock and Grossberg, 1988), including data on synchronous movement completion by different joints (Freund and Büdingen, 1978), muscle contraction duration invariance (Freund and Büdingen, 1978; Ghez and Vicario, 1978), bell-shaped velocity profiles (Howarth and Beggs, 1971), changing velocity profile asymmetry at higher movement speeds (Beggs and Howarth, 1972; Zelaznik, Schmidt, and Gielen, 1986), amplification of peak velocity during target switching (Georgopoulos, Kalaska, and Massey, 1981), and speed-accuracy tradeoffs (Fitts, 1954; Woodworth, 1899). Furthermore, Guenther (1992) and Bullock et al. (1993) showed that directional tuning curve properties of neurons utilized in such a mechanism closely match the properties of cells found in monkey motor cortex (e.g., Georgopoulos, Kalaska, Caminiti, and Massey, 1982; Caminiti, Johnson, and Urbano, 1990).

The amplitude strategy can be carried out in DIVA by changing the size of the convex region target, as shown in Figure 10. Here, the orosensory target used to produce a particular sound at a slow speaking rate is formed by "shrinking" the convex region learned during babbling for that sound. This can be interpreted as a tendency for speakers to hyperarticulate, or use a more "canonical" configuration of the vocal tract, when producing a phoneme at slower rates, leading to clearer, more precise speech when rate constraints are less stringent (e.g.,Lindblom, 1963; 1983; 1990). The use of hyperarticulation for other purposes is discussed in Section 9.

The act of increasing convex region size for increased movement speeds is very much in the spirit of the well-documented speed-accuracy trade-off of movement control described by Fitts' Law (see Schmidt, 1982, for a review). Fitts (1954) showed that for back and forth targeted arm movements of a fixed distance, increasing the size of the targets allowed subjects to increase movement speeds. This relationship has been shown to hold for many other movement tasks, including arm movements to a single target (Fitts and Peterson, 1964), wrist rotations (Knight and Dagnall, 1967), and head movements (Jagacinski and Monk, 1985). Increasing the size of the convex region target during faster

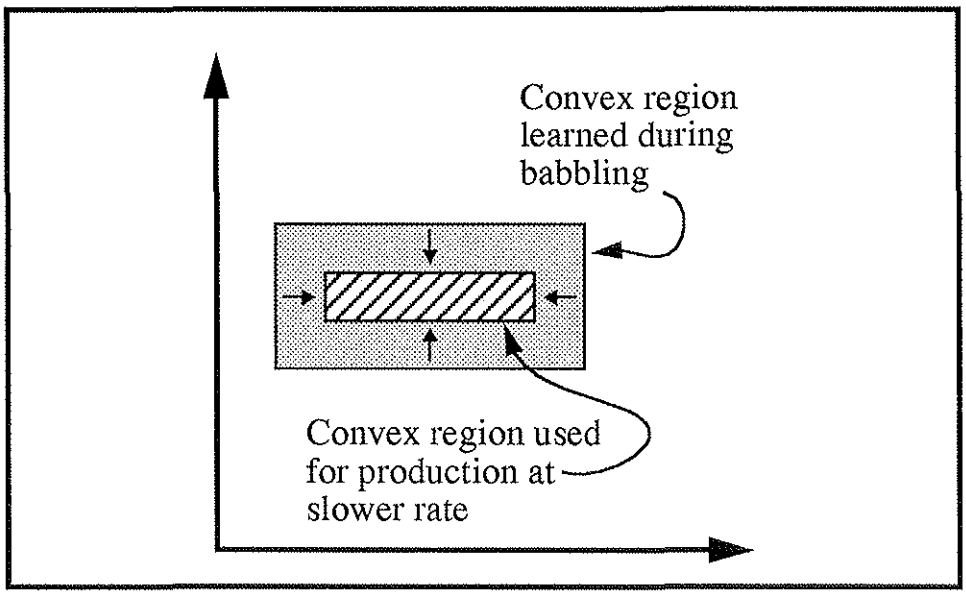

FIGURE 10. The amplitude strategy of changing speaking rate can be carried out in DIVA by shrinking the convex region target used for production at slower speaking rates. This corresponds to using a more "canonical" target position for increased clarity at slower speaking rates.

speech in DIVA is likewise a case of trading off accuracy for speed, this time in the domain of speech production. The concept of a target as a convex region whose size can be varied -- rather than a single point as is typically assumed in models of movement control -- seems naturally suited for explaining how subjects adjust the accuracy of movements when speed requirements are increased.

Shrinking of the convex region target for a sound can be carried out surprisingly easily in the DIVA neural network: simply add a small positive input to all ODV cells. Since the same input is added to all ODV cells, this input will be referred to as *non-specific*. To see why a non-specific input shrinks the convex region targets, consider a single antagonistic pair of ODV cells, corresponding to a single dimension of orosensory space. Since only positive ODV activities can drive movement, the size of the convex region target for a sound along that dimension corresponds to the range of values of orosensory feedback that result in no positive activity of either ODV cell in the antagonistic pair (see Section 3). If a positive non-specific input is added to both cells in the pair, the range of orosensory feedback values that result in no positive activity of either ODV cell, and thus the size of the convex region target, is reduced. If the same non-specific input is added to all ODV cells,

the convex region target shrinks toward the center along all dimensions as schematized in Figure 10. Because a larger tonic activity results in a smaller, more precise target, the size of this input should be inversely related to movement speed. To achieve this, we can modify Equations 3 and 4, which govern ODV cell activity, as follows:

$$d_{j+} = \sum_i s_i z_{ij+} - f_{j+} + R\,(1 - G) \tag{16}$$

$$d_{j-} = \sum_i s_i z_{ij-} - f_{j-} + R\,(1 - G) \tag{17}$$

where G is the value of the GO signal (varying between 0 and 1), and R is a parameter that corresponds to the degree to which a particular speaker prefers the amplitude strategy to the velocity strategy. The non-specific input to the ODV cells is thus $R\,(1 - G)$, which varies inversely with volitional movement speed as embodied by the GO signal activity G. Adding a positive input to both channels in an ODV antagonistic pair can have an undesireable side-effect: it can result in positive activities at both ODV cells in the pair. Conceptually, this is like commanding both an increase and a decrease of an orosensory variable such as lip aperture. This problem is easily avoided by changing Equations 7 and 8, governing AVV activity during performance, as follows:

$$a_{k+} = \sum_j \left[[d_{j+}]^+ - [d_{j-}]^+ \right]^+ w_{j+k+} + \sum_j \left[[d_{j-}]^+ - [d_{j+}]^+ \right]^+ w_{j-k+} \tag{18}$$

$$a_{k-} = \sum_j \left[[d_{j+}]^+ - [d_{j-}]^+ \right]^+ w_{j+k-} + \sum_j \left[[d_{j-}]^+ - [d_{j+}]^+ \right]^+ w_{j-k-} \; . \tag{19}$$

These equations imply a competitive interaction between antagonistically paired cells in the ODV stage.

In the DIVA simulations, the velocity and amplitude strategies are used simultaneously to increase speaking rate. However, the two strategies are used to different degrees in different simulations to account for the various speakers seen in the data mentioned above. This is accomplished in the model by changing the parameter R. A value of R close to 0.0 simulates a speaker who relies more on the velocity strategy than the amplitude strategy, while a larger value of R simulates a speaker who relies more on the amplitude strategy.

Table 4 shows simulation results of the model producing the phrase /pap/ (cf. Gay et al., 1974; Kuehn and Moll, 1976) at two different speeds and using two different values of the R parameter. The maximum velocities of the gestures used to produce the speech sounds (tongue body movements for the vowel, lower lip movements for the consonant) are given in the first two rows of each table. With R = 0.0, the model preferentially uses the velocity strategy. This is an extreme case where the amplitude strategy is completely unused. Here, we can see that maximum velocities of movements toward both vowels and consonants increase (top two rows of the top half of Table 4). This is in concert with data from Kuehn and Moll (1976) and Ostry and Munhall (1985) for subjects who rely on the velocity strategy.

Much more interesting is the case where the model preferentially uses the amplitude strategy with R = 0.2. Despite the fact that the model uses the same strategy to produce vowels and consonants, vowel movement velocities *decrease* with increased speaking rates, while consonant velocities increase (top two rows of the bottom half of Table 4). This is precisely the behavior reported by Gay et al. (1974), Kuehn and Moll (1976), and Ostry and Munhall (1985) for subjects using the amplitude strategy, and this is the result used as evidence for different control structures for vowels and consonants by Fowler (1980).

TABLE 4. Simulation Results Showing Effects of Speaking Rate on Vowel and Consonant Movement Kinematics During the Utterance /pap/

Quantity measured	Slow rate (G = 0.5)	Fast rate (G = 1.0)	Percent change
VELOCITY STRATEGY (R = 0.00)			
Vowel maximum velocity	0.010	0.019	+90%
Consonant maximum velocity	0.088	0.176	+100%
Vowel max. velocity / distance	0.011	0.020	+81%
Consonant max. velocity / distance	0.020	0.040	+100%
AMPLITUDE STRATEGY (R = 0.20)			
Vowel maximum velocity	0.034	0.019	-44%
Consonant maximum velocity	0.144	0.176	+22%
Vowel max. velocity / distance	0.010	0.020	+100%
Consonant max. velocity / distance	0.025	0.040	+60%

Why do vowels and consonants show such different behavior despite being treated exactly the same in the model? The answer lies in the nature of the convex regions learned during babbling for the two sound types. Figure 11 schematizes this situation. Even along important orosensory dimensions such as tongue body position with respect to the maxilla, acceptable vowels can be produced within a relatively large range of positions. Consonants, on the other hand, require very strict control along important orosensory dimensions to insure either full closure (for stops) or frication (for fricatives). During babbling, therefore, the model learns convex regions reflecting these properties, as shown in Figure 11. Now consider what happens when the two convex region types are shrunk toward their centers for slower speech according to Equations 16 and 17. The distance the vocal tract must move to reach the target during slow speech and fast speech are labeled D_S and D_F , respectively. For a given initial vocal tract configuration (represented by the black dot in the figure), shrinking the convex region for a vowel results in a much bigger change in the distance needed to travel to the target (D_S - D_F) than shrinking the convex region for a consonant. This tendency for vowel movements to show decreased displacements is commonplace in human speech and is termed *vowel reduction* (e.g., Lindblom, 1963; 1983). Furthermore, earlier results showed that movement speed was directly related to movement distance in DIVA (see Section 5). Because movement distance decreases much more for vowels than consonants at fast rates as compared to slow rates, we see a much smaller velocity increase, or even a decrease, for vowels spoken at a fast rate when using the amplitude strategy. (Note that both vowels and consonants also receive a larger GO signal value G in Equation 9 at faster rates; this is why the consonant movement speed increases despite little or no change in movement distance. In vowels, the increase in G is more than offset by the decrease in movement amplitude, which is reflected in decreased activities a_{k+} and a_{k-} in Equation 9.)

A second telling aspect of the simulation results shown in Table 4 is revealed by looking at the ratios of maximum velocity to movement distance for the vowel and consonant gestures (bottom two rows of the top and bottom halves of Table 4). This ratio increases with increased speaking rate for both vowels and consonants, *regardless of strategy used.* This is rather surprising in the case of vowel movements under the amplitude strategy; even though maximum velocity decreases, the ratio of maximum velocity to movement distance increases. This phenomenon was the central focus of the study by Ostry and Munhall (1985). These investigators found that speakers showed an increase in V_{max}/distance

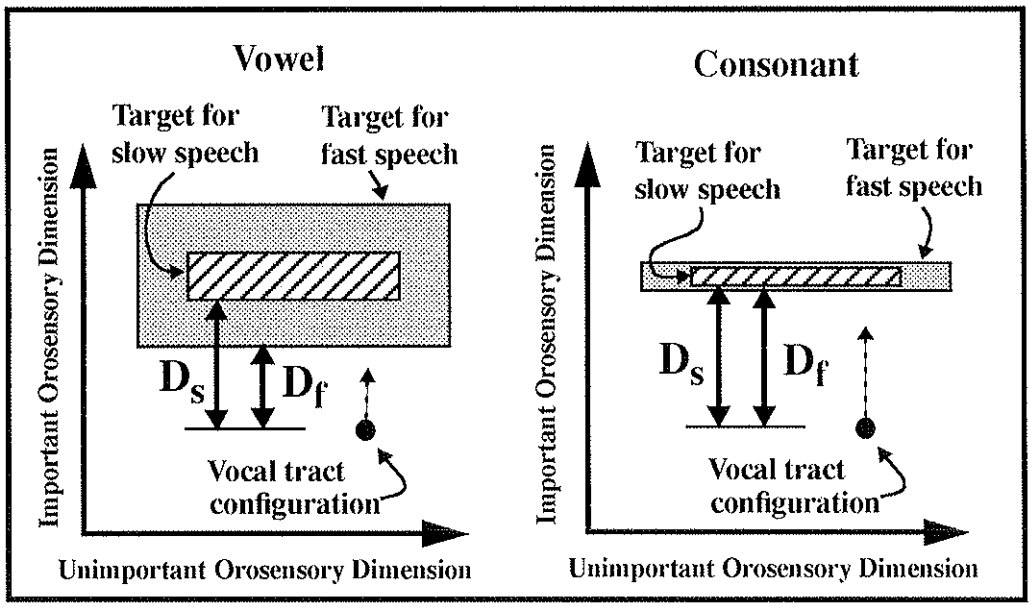

FIGURE 11. Differential effects of convex region shrinkage for vowels and consonants. The convex regions learned for vowels during babbling (left) allow for larger variability along important orosensory dimensions than the convex regions for consonants (right). This is because consonants require an essentially invariant constriction of the vocal tract for production. For a given configuration of the vocal tract (black dot), shrinking the convex region for slower production of a vowel results in a larger change in the distance of movement required to get to the target (D_s - D_f) than shrinking the convex region target for a consonant. This results in the differential speaking rate effects seen in the articulator movements for vowels and

independent of whether they favored a velocity strategy or an amplitude strategy, as seen in the simulations summarized in Table 4.

In DIVA, this results from the multiplicative interaction between the GO signal and movement distance, described by Equation 9. Rearranging this equation yields the following:

$$\frac{v_k}{[a_{k+} - a_{k-}]} = G \tag{20}$$

where v_k is the velocity along the k^{th} articulatory degree of freedom, G is the value of the GO signal, and a_{k+} and a_{k-} are the antagonistically paired AVV activities corresponding to the i^{th} articulatory degree of freedom. Through the dynamics of Equations 16-19, the size of the activities a_{k+} and a_{k-} reflect the distance to the current target. Equation 20 thus indicates that the velocity/distance ratio scales with the GO signal. Since increases in speaking rate are carried out though increases in the GO signal, it is clear that the ratio of

velocity to distance will increase at faster speaking rates. (This can also be seen as the increase in slope for faster speaking rates in the plot of maximum velocity vs. distance for vowel movements in Figure 7.) In the case of vowel movements using an amplitude strategy, the decrease in articulator velocity at faster speaking rates occurs because of an even larger decrease in movement distance that outweighs the increase in G of Equation 9; thus, the ratio of peak velocity to distance increases despite a decrease in peak velocity.

Finally, it is interesting to note that the displacements of jaw movements for both vowels and consonants decrease with increased speaking rate in DIVA simulations using the amplitude strategy. This result was reported for human subjects in the study of Gay et al. (1974). Table 5 illustrates this for vowel and consonant gestures produced by the model in the phrase /apapapa/. This phenomenon disappears as the parameter R is decreased (i.e., as a shift to the velocity strategy is implemented), with the movement displacements being equal at fast and slow rates when R is set to 0.0. This is of interest because some subjects have shown little effect of speaking rate on jaw displacement (e.g., Abbs, 1973). The simulation results reported here suggest that these subjects may have used the velocity strategy, whereas the Gay et al. (1974) subjects were known to have used the amplitude strategy.

TABLE 5. Simulation Results Showing Jaw Displacement During the Utterance /apapapa/ Using the Amplitude Strategy

	Average Jaw Displacement	
Gesture type	Slow rate (G = 0.5)	Fast rate (G = 1.0)
Vowel	48.7	37.8
Consonant	43.4	34.1

This section has shown how the convex region theory, generated to explain how infants can learn acceptable limits of variability for articulator movements, provides an insightful and parsimonious explanation of a collection of speaking rate effects not previously treated by a single model. This explanation arises from two basic mechanisms in the model, both of which are supported by ample psychophysical data. The first, a multiplicative GO signal, was originally posited by Bullock and Grossberg (1988) to explain a wide range of data on arm movements. Furthermore, the increase in V_{max}/distance with increased speaking rate reported by Ostry and Munhall (1985), even when subjects pro-

duced slower movement velocities for vowels at faster rates, directly implicates such a mechanism. The second mechanism, a non-specific input to ODV cells that shrinks the size of convex region targets for slower speaking rates (thereby increasing movement amplitudes at slower rates, particularly for vowels), is implicated by data on vowel reduction and captures the essence of the speed-accuracy trade-off described by Fitts' Law. The model's explanation asumes no differences in strategy for vowels and consonants, yet differential effects arise for the two sound types with changes in speaking rate. Finally, individual differences in strategies across speakers are captured by variation of a single parameter R.

8. Carryover Coarticulation

We now address data on carryover coarticulation, also known as perseveratory or left-to-right coarticulation. Carryover coarticulation refers to cases when the vocal tract configuration for one segment influences the configuration or sound for a later segment. Carryover coarticulation most likely covers several distinct phenomena, as posited in the following paragraphs.

One form of carryover coarticulation results from the fact that movements to and from a speech segment follow different paths depending on context (e.g., Daniloff, Schuckers, and Feth, 1980). For example, when producing the syllables /at/ and /it/, the paths taken by the articulators to reach /t/ differ because of different starting configurations from the preceding vowels. This form of carryover coarticulation results in DIVA because the Orosensory Direction Vector activities, which drive movement of the articulators, depend on the current configuration of the vocal tract. Simply stated, the model moves in an approximately "straight line" trajectory from the current configuration to the target configuration.

A more interesting case of carryover coarticulation in DIVA occurs because the configuration of the vocal tract when movement starts toward a segment's target determines where on the convex region the vocal tract ends up. This is schematized in Figure 12a for the target /k/ in the words "luke" and "leak". Here, the initial front-back position of the tongue body for the preceding vowel determines the configuration of the vocal tract reached for the consonant /k/. When the back vowel /u/ precedes /k/ as in "luke", the tongue body is further back during /k/ than when the front vowel /i/ precedes /k/ as in "leak".

A simulation verifying this property is shown in Figure 13a. The "+" marks front-back position of the stop for "luke". Comparison of the stop location during "leak" reveals the anteroposterior variation reported for human subjects when producing these words (e.g., Daniloff et al., 1980; Kent and Minifie, 1977). As schematized in Figure 12a, variability results in DIVA because the vocal tract configuration for /k/ moves to the closest point on the convex region target; thus, the model reproduces the "economy of effort" seen in human speech (Lindblom, 1983) by moving from the vocal tract configuration for the vowel to the closest acceptable configuration for the sound /k/.

A final case of carryover coarticulation occurs when one aspect of a segment's configuration is maintained for one or more following segments. For example, lip protrusion for the /u/ in "spoon" is maintained through the /n/ (Daniloff and Moll, 1968). In DIVA, this occurs automatically when the position of the vocal tract for the preceding sound (the /u/ in this case) along the orosensory dimension in question lies within the convex region of the target for the following sound (the /n/ in this case) along the same dimension. This is schematized in Figure 12b, and a simulation result showing carryover coarticulation of lip protrusion for "spoon" is shown in Figure 13b.

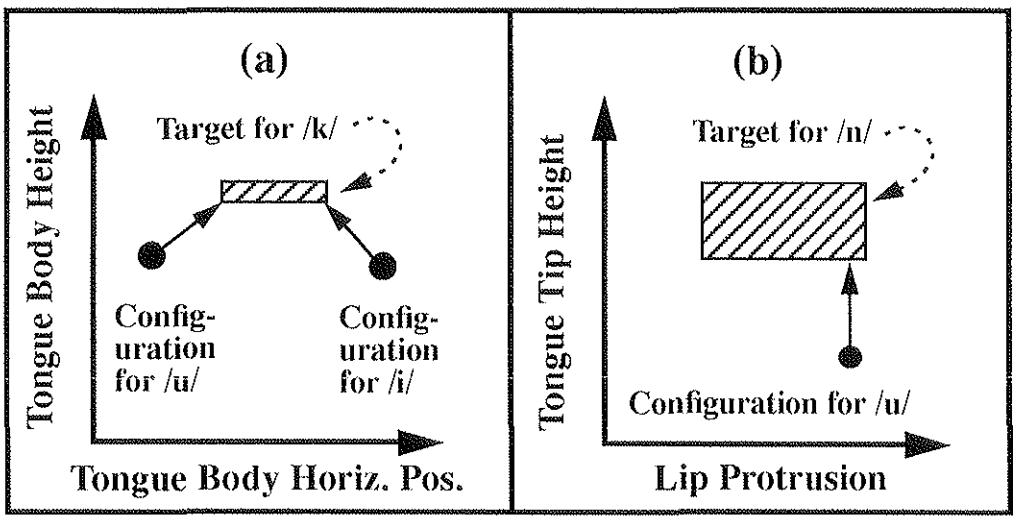

FIGURE 12. Two schematized cases of carryover coarticulation in DIVA. (a) Approaching the target for /k/ from the configuration corresponding to the back vowel /u/ in "luke" leads to a final tongue body configuration that is further back than when approaching from the configuration corresponding to the front vowel /i/ in "leak". (b) When moving from the configuration reached for /u/ (filled circle) to the target for /n/ in "spoon", position along the Lip Protrusion dimension is already within the convex region for /n/ along that dimension, so the lips are not retracted.

It should be noted that in this case carryover coarticulation is the result of a general tendency not to move an articulator unless it needs to be moved. In the /n/ of "spoon", the protruded lips did not need to be retracted since they already fell within the convex region target for the following /n/. In a recent X-ray motion film study of articulator movements, Wood (1991) notes that instances of perseveratory coarticulation seen in his data "all seem to be examples of the ... tendency for individual articulators to be left idle until required again" (p. 290). As an example, Wood points out that lip protrusion for a rounded vowel is retracted only slightly during the following stop, then retracted fully only for the unrounded vowel following this stop. This phenomenon is shown in DIVA for the phrase /udi/ in Figure 13c. In this figure, the "+" denotes the lip protrusion position during the /d/ for purposes of comparison. Clearly, the lips are fully extended for /u/, are only partially retracted for the stop /d/, and are fully retracted only when required for the unrounded vowel /i/, as seen in the Wood data.

In contrast, the task-dynamic model of Saltzman and Munhall (1989) utilizes a "neutral attractor" that moves unused articulators toward a neutral configuration. One reason for this attractor is investigated in the next section. In the task-dynamic model, one would thus expect unused articulators to be constantly moving unless they were already in the neutral configuration; this is not compatible with the data from Wood (1991), however, where unused articulators remained stationary even if they are not in a neutral configuration, e.g. when the lips are protruded.

It is often hypothesized that carryover coarticulation results largely from mechanical or inertial effects involved in moving the articulators from one sound's target to the next rather than from explicit pre-planning as seen in anticipatory coarticulation (e.g., Baum and Waldstein, 1991; Daniloff et al., 1980; Flege, 1988b; Gay, 1977; Recasens, 1987, 1989). However, as pointed out by Daniloff and Hammarberg (1973), the mechano-inertial explanation is inadequate since large carryover effects are seen at low speeds and may spread over two or three segments, indicating a deliberate process for producing these effects. Based on a study requiring subjects to begin an utterance before knowing its end, Whalen (1990) also hypothesized that carryover effects are probably largely planned, but to a lesser degree than anticipatory effects.

It is interesting to note that carryover coarticulation in DIVA results solely from the dynamics of moving between targets and not from an explicit pre-planning mechanism

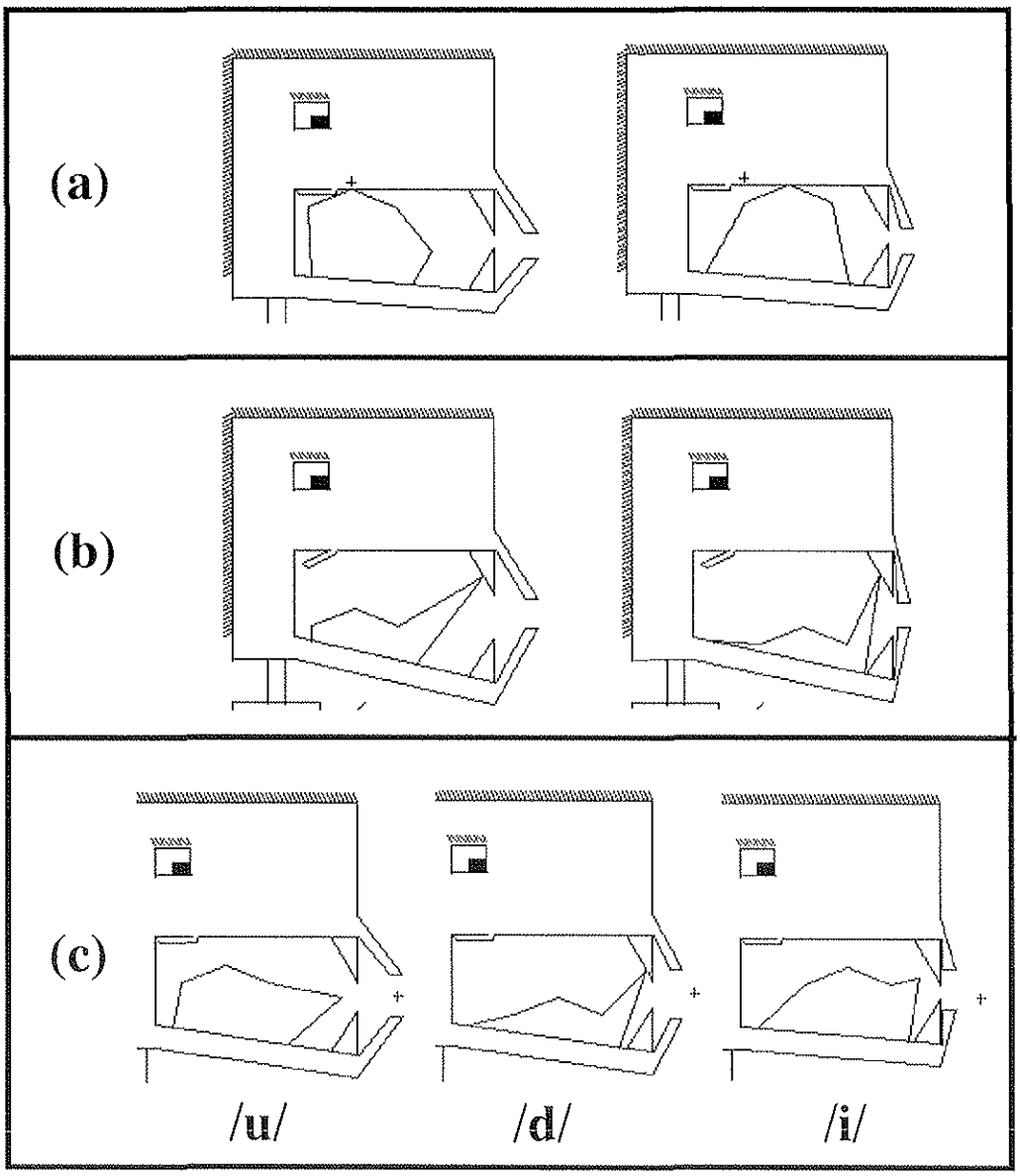

FIGURE 13. Simulations of carryover coarticulation in the DIVA model. (a) Coarticulation of tongue body constriction place for the velar stop /k/ in "luke" (left side) and "leak" (right side). The "+" marks the tongue body constriction location during the /k/ in "luke" for comparison. This simulation also shows carryover coarticulation of lip protrusion for "luke".(b) Coarticulation of lip protrusion during the /n/ of "spoon" (left side). The right side shows the configuration for /n/ in "span" for comparison. (c) Coarticulation of lip protrusion in the utterance /ude/ as seen in the data of Wood (1991). The "+" marks position of the lips during /u/ for comparison. The lips are fully extended for /u/ (left side), then only partially retracted for /d/ (center) before being fully retracted only when required for /i/ (right side).

(cf. the explanation of anticipatory coarticulation in the next section). Nonetheless, these effects are not mechano-inertial; in fact, the articulators are treated as weightless. Instead, they are "planned" in the sense that they result from explicit movement commands from the production mechanism. This planning does not require advance knowledge of later segments, but instead arises from the interaction between the configuration of the vocal tract at the start of a segment and the convex region target for the segment. Carryover coarticulation can continue over several segments, however, if the vocal tract configuration along a particular orosensory dimension at the start of the segments lies within the convex region targets of these segments along that dimension (see Figure 12b). The DIVA explanation of carryover coarticulation thus accounts for the seemingly incongruous observations that carryover coarticulation can occur with knowledge only of the next segment to be produced (as suggested by the results of Whalen, 1990), yet carryover effects can extend for several segments (as pointed out by Daniloff and Hammarberg, 1973).

9. Anticipatory Coarticulation

Based on the pioneering work of researchers such as Kozhevnikov and Chistovich (1965), Henke (1966), and Öhman (1966), the literature on anticipatory, or right-to-left, coarticulation has been dominated by two categories of models: *look-ahead models* and *coproduction models* (for recent comparisons, see Boyce et al., 1990; Fowler and Saltzman, 1993; Whalen, 1990; Wood, 1991). This section first briefly describes these two model types, then identifies a common shortcoming of the two concerning the nature of phoneme targets commonly assumed in both models. A generalization of the look-ahead model based on convex region targets is then defined. Finally, an implementation of the generalized look-ahead approach in the DIVA model is compared to coarticulation data.

The look-ahead model of anticipatory coarticulation (e.g., Henke, 1966; Kozhevnikov and Chistovich, 1965; Perkell, 1980), considered here to include the closely-related feature spreading model (e.g., Daniloff and Hammarberg, 1973), is best understood by considering a phoneme as a bundle of "features" (Chomsky and Halle, 1968; Jakobson and Halle, 1956), each describing the configuration of only a small portion of the vocal tract. Each phoneme uses a subset of the possible features. The model explains coarticulation by positing that movements for a feature of a later segment can start as long as the current segment and any intervening segments do not use that feature. For example, in a /vcccv/

sequence where the final vowel is rounded but none of the preceding sounds use that feature, production of the feature "round" can begin as early as the first vowel. This was in fact reported for human subjects in a study by Benguerel and Cowan (1974), although disputed elsewhere (Boyce et al., 1990).

In the coproduction model (e.g., Öhman, 1966, 1967; Fowler, 1980; Saltzman and Munhall, 1989), vowel and consonant gestures have fixed time courses, but these time courses can be overlapped in time with the time courses of neighboring gestures. Öhman (1966, 1967) hypothesized that this is possible because vowels and consonants use largely independent subsets of the vocal tract musculature. Fowler (1980) repeated this sentiment, hypothesizing that different coordinative structures exist for the two sound types. This idea has been further refined in the work of Saltzman and Munhall (1989), who use a set of "blending parameters" that govern the relative effects of the different coordinative structures in cases where two or more simultaneously active coordinative structures involve the same musculature. Within a coproduction framework, coarticulation arises simply because vowels and consonants can be overlapped in time, or "coproduced". In a coproduction model, the target time courses for segments are the same regardless of context, whereas in the look-ahead model the time course of a segment can be changed by starting production of one of its features earlier in time when possible. In the example of /vcccv/ sequences with a final rounded vowel, the coproduction model predicts that the beginning of lip rounding for the final vowel will be time-locked to the acoustic onset of the vowel; this was reported by Bell-Berti and Harris (1979), seemingly contradicting the results of Benguerel and Cowan (1974) mentioned above. In fact, much supporting data has been posited for both theories. Recent attempts have been made to reconcile much of these data with a coproduction model (Boyce et al., 1990; Fowler and Saltzman, 1993), but other recent work claims more experimental support for the look-ahead model (e.g., Wood, 1991). In short, the debate over the two model types continues nearly 30 years after publication of their theoretical roots.

It is useful to investigate the nature of speech targets typically assumed in the two theories. Figure 14a schematizes the typical form of targets in both look-ahead and coproduction theories of coarticulation. Both of these theories posit that each sound utilizes only a subset of the vocal tract. For example, vowels specify tongue body height but not velum height, and bilabial consonants specify lip aperture but not tongue body height. Thus, if

we look at the target for a vowel along orosensory dimensions corresponding to velum height and tongue body height, we will see that a strict target position of the tongue body is specified, but velum height is totally unspecified, as shown in Figure 14a.

In contrast, the convex region theory posits a vowel target as shown in Figure 14b. Here, a small range of tongue body positions are included in the target, and a large but not complete range of velum heights are included. Instead of the "all or nothing" nature of traditional targets, wherein each orosensory dimension is either strictly specified or not specified at all, a convex region target specifies target ranges for all orosensory dimensions, with the size of the ranges varying from very small (e.g., in the case of lip aperture for bilabial consonants) to very large (e.g., in the case of tongue body height for bilabial consonants). Traditional targets can thus be thought of as a special case of convex region targets, formed by "binarizing" the size of the target range along each orosensory dimension.

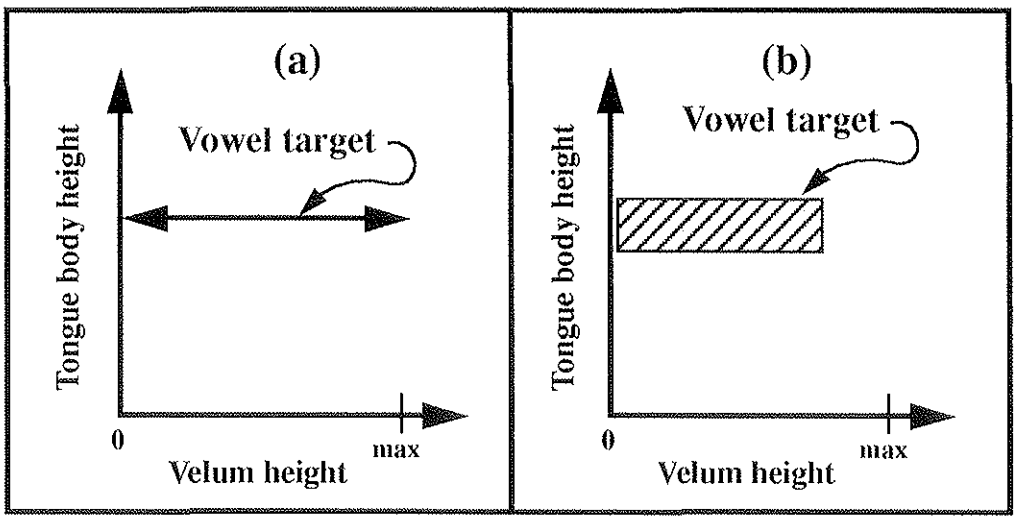

FIGURE 14. (a) Typical vowel target assumed within the coproduction and look-ahead theories of anticipatory coarticulation. The target specifies a tongue body height, but velum height is completely unspecified, thus potentially allowing a velum position anywhere within the entire range from completely closed to maximally opened. (b) The analogous target within the convex region theory. This target specifies a small range of tongue body heights and a large range of velum heights. Note, however, that *some* limits are placed on velum height; for example, the velum is not allowed to be completely open.

Considerable evidence favoring targets of the form shown in Figure 14b has already been given in this article. For example, the data from Kent et al. (1974) discussed and simulated

in Section 6 indicate that although velum height can vary widely for a vowel, it is not completely unspecified as in traditional targets (see Figure 9). Fowler and Saltzman (1993, p. 187; see also Bell-Berti, 1980) also point out that vowels have some target specification of velum height, but no velum target appears to be used for vowels in the Haskins linguistic gestural model and task dynamic model (e.g., see Browman and Goldstein, 1990, p. 345). Such underspecification of the vocal tract can lead to problems when mapped into articulator movements. For example, vowel gestures in the Haskins models do not include a target value for lip aperture (e.g., Saltzman and Munhall, 1989, p. 343). If no corrective mechanism is added, the lips would remain closed during a vowel between two bilabial stops, e.g. when producing the word "bob". Of course, this would not result in proper production of the vowel. The problem is overcome in the task-dynamic model by incorporating a "neutral attractor", which acts as a default target when no other target value is specified. Implementing this requires the addition of several matrix terms to an already complex dynamical system, including a gating matrix specifically designed to prevent the neutral attractor from interfering with actively commanded movements. (Recall also the evidence from Wood, 1991, against neutral attractor effects on unused articulators described in Section 8.) The neutral attractor amounts to a supplemental target needed to overcome underspecification of the vocal tract that results from using all or nothing targets as schematized in Figure 14a.

It thus appears that the all or nothing nature of traditional targets is a simplification that may belie the true nature of phonemic targets, which involve much more of the vocal tract than is typically assumed. Convex region targets, on the other hand, do not underspecify the shape of the vocal tract. Instead, they specify exactly the range of variation allowable along every orosensory dimension. Along the dimension of lip aperture for a vowel target, the target range does not include complete closure, so the problem described above when producing "bob" does not arise. This explanation, wherein all of the vocal tract requirements of a phoneme are encoded in its target, seems much more natural than using additional machinery (e.g., a neutral attractor) to prevent accidental violations of the vocal tract requirements for a phoneme. Furthermore, because convex region targets specify a range of target values rather than a point target for each orosensory dimension, the potential problem of *overspecifying* vocal tract shape, and thus commanding unnecessary articulator movements, is also avoided.

Because each convex region target is defined over all orosensory dimensions, the current model would appear at first glance to be incompatible with a look-ahead model of coarticulation since the latter requires some dimensions to be unused in the current phoneme's target so that features from future phonemes can spread back in time. However, much as the convex region target represents a generalization of the traditional target (Figure 14), we can define a generalized version of the look-ahead model that replaces "binary" concepts with more continuous concepts. The key idea behind this model is schematized in Figure 15 for the word "coo". The target for /k/ and the target for /u/ overlap along the orosensory dimension of lip protrusion. Therefore, when producing /k/, we can use a reduced target for /k/ that only includes the region of overlap along the orosensory dimension of lip protrusion. This "coarticulated" target is outlined by the bold rectangle in Figure 15. If we move the vocal tract to the coarticulated target, we will see anticipatory lip protrusion for /u/ during production of /k/.

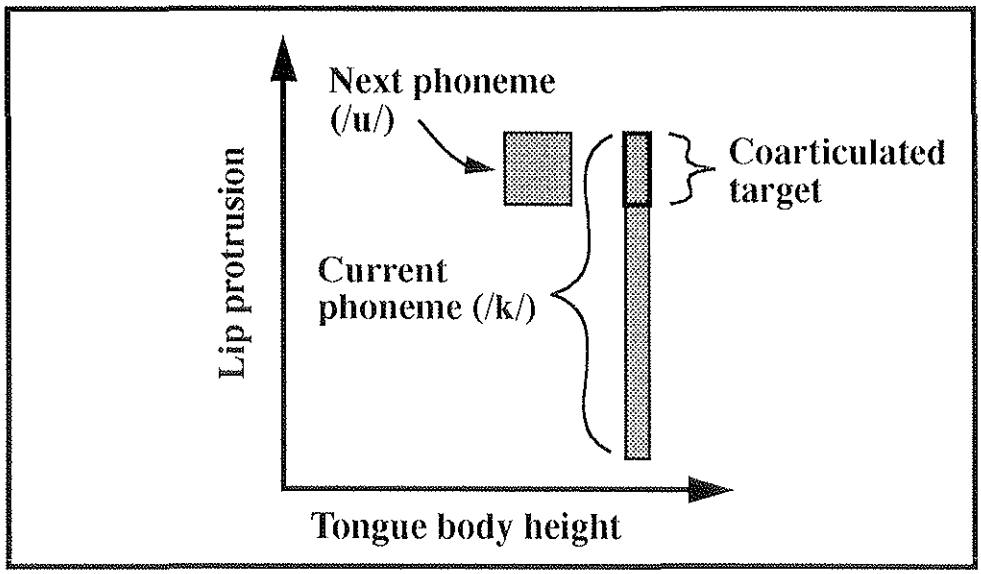

FIGURE 15. Generalization of the look-ahead model implemented by DIVA, schematized for the word "coo". The convex region targets for /k/ and /u/ overlap along the dimension of lip protrusion. When pronouncing /k/, the target is shrunk to include only the overlapping portion along the lip protrusion dimension. Since no overlap occurs for the tongue body height dimension, the full target range for /k/ is used. Movements to the "coarticulated" target will thus lead to anticipatory lip protrusion during /k/, as seen in human speakers and verified in the simulation results of Figure 16.

The generalized look-ahead model can be stated more precisely as follows:

> For each orosensory dimension, the coarticulated target starts out as the target range of the current phoneme along this dimension. If the coarticulated target and the target of the next phoneme overlap along this dimension, the coarticulated target is reduced in size to the region of overlap, and the process is repeated for the next phoneme in the string. If there is no overlap, no further look-ahead is performed along this dimension.

Note that if we replace the notion of a feature with the notion of an orosensory dimension, the traditional look-ahead model can be seen as a special case of the above in which target ranges for a phoneme are either a single point (when the feature is specified for the phoneme) or the entire possible range along the orosensory dimension (when the feature is unused for the phoneme).

The generalized look-ahead approach is currently implemented algorithmically, rather than by explicit neural network circuitry, in DIVA. All simulations reported in this article used this look-ahead procedure. As is the case with all versions of the look-ahead model, this algorithmic process implicitly assumes that future phonemes exist in a memory buffer, and that the current phoneme's target can be affected by the targets for these future phonemes as described above. The DIVA model is capable of looking ahead an arbitrary number of phonemes, but the simulation results reported here use a procedure limited to a look-ahead window of two phonemes. It seems likely that if humans indeed use a look-ahead process, then they are capable of varying the size of the look-ahead window, perhaps as a function of the number of phonemes in the memory buffer. Such utterance-specific variability might explain why speech experimentalists have been unable to convincingly demonstrate whether the coproduction model or look-ahead model better describes human anticipatory coarticulation. It should also be noted that the generalized look-ahead process is not inherent to DIVA per se. In fact, a generalized coproduction model could similarly be defined using convex region targets and implemented in the DIVA architecture, and future research will likely compare the properties of such an implementation with the generalized look-ahead implemenation.

Simulation results showing anticipatory coarticulation in DIVA are shown in Figure 16. The top row of this figure shows the configurations reached by the model during production of /k/ in the words "coo" (left side) and "key" (right side). Two forms of anticipatory coarticulation can be seen. First, lip rounding in anticipation of the rounded vowel /u/ is

seen for "coo", with no anticipatory lip rounding for the unrounded vowel /e/ in "key", as seen in human data (e.g., Benguerel and Cowan, 1974). Second, the horizontal location of the velar constriction for /k/ is further back in anticipation of the back vowel /u/ in "coo", and further forward in anticipation of the front vowel /i/ in "key", again as seen in human speech (e.g., Daniloff et al., 1980, p. 328).

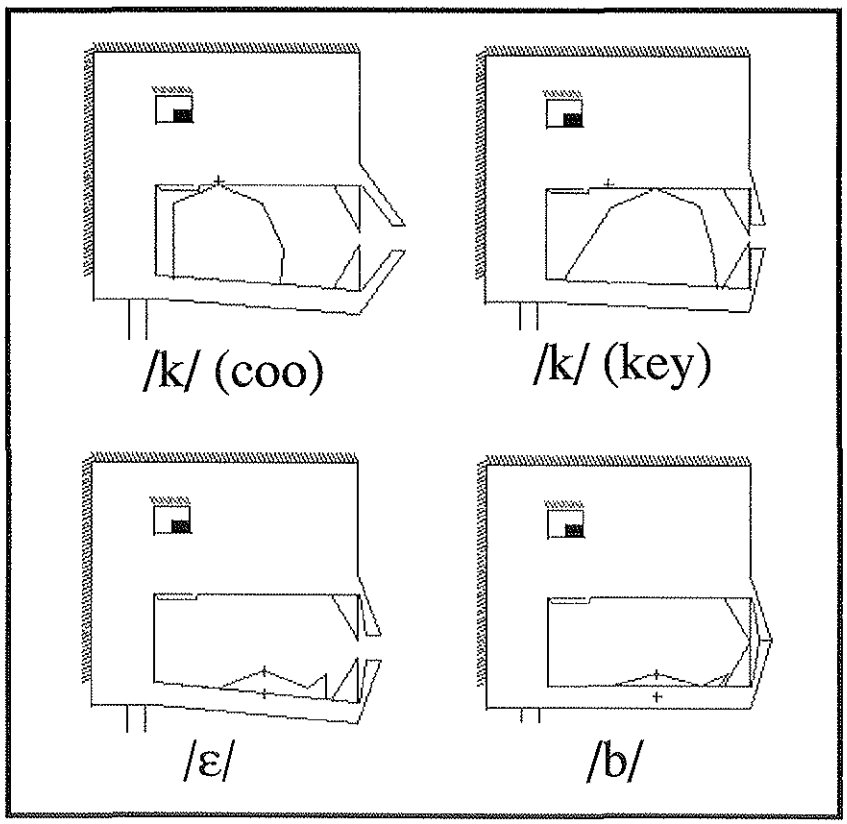

FIGURE 16. Simulation results showing anticipatory coarticulation. "+" marks are for purposes of comparison. Top row shows configuration during production of /k/ in the words "coo" (left side) and "key" (right side). Anticipatory lip rounding for /u/ is seen for "coo", but no anticipatory rounding is seen for the unrounded vowel /i/ in "key". Furthermore, the place of velar constriction is toward the back for "coo" in anticipation of the back vowel /u/, and toward the front in "key" in anticipation of the front vowel /i/. Bottom row shows the configurations during production of /ɛ/ and /b/ for the utterance /ɛbɛ/. The tongue body is depressed slightly with respect to the jaw during /b/ in anticipation of the mid vowel /ɛ/ as seen in the study of Wood (1991).

The bottom row of Figure 16 shows a much more subtle situation where anticipatory coarticulation and motor equivalent compensation are used in concert. Wood (1991, p. 290)

describes data from a subject producing the biabial stop /b/ between two mid vowels. If the tongue body remained at the same position with respect to the jaw while the jaw raised for /b/, the vocal tract configuration would have been moved toward the configuration for the high vowel /i/ (i.e., tongue body height with respect to the maxilla would increase). Instead what was seen was a compensatory lowering of the tongue body position with respect to the jaw, allowing the subject to maintain the mid tongue body configuration in anticipation of the following mid vowel. The bottom row of Figure 17 shows results from a DIVA simulation of the utterance /ɛbɛ/. The left figure shows the configuration of the vocal tract for the first /ɛ/ in the utterance. The "+" marks identify the height of the tongue body and jaw for comparison. The right side shows the configuration during /b/. This figure clearly shows that the tongue body height with respect to the maxilla is maintained for the ensuing mid vowel /ɛ/ despite raising of the jaw; this happens through a compensatory lowering of the tongue body with respect to the jaw. The model is anticipating the mid height along the important orosensory dimension of tongue body height with respect to the maxilla, and efficient use of the redundant articulator system as described in Sections 3 and 4 allows the model to automatically compensate for jaw raising. Such an effect would be impossible to predict with models that investigate coarticulation in isolation of other speech production competencies such as motor equivalence.

The generalized look-ahead model described here avoids a problem pointed out by Fowler and Saltzman (1993) for look-ahead models that use the types of targets schematized in Figure 14a. These authors note that look-ahead models cannot predict transconsonantal vowel-to-vowel anticipatory coarticulation in /vcv/ sequences because all vowels utilize the same features, specifying the vowel's height, frontness, and lip configuration. Such vowel-to-vowel coarticulation has been reported for human subjects (e.g., Kent and Moll, 1972b; Manuel, 1990; Öhman, 1966). The problem arises in the traditional look-ahead model because presence of a feature in the current phoneme precludes spreading of that feature from future phonemes. In the generalized look-ahead model, target ranges replace this all or nothing notion of a feature. As long as the ranges for the first vowel, consonant, and second vowel overlap along any orosensory dimension, vowel-to-vowel coarticulation will be seen.

Because the configuration used to produce a phoneme cannot extend beyond its convex region target, the amount of coarticulation produced by the generalized look-ahead model

depends very much on the size of the convex regions. Smaller targets will necessarily reduce the amount of coarticulation that can arise. Similarly, Manuel (1987; 1990) hypothesized that languages with more crowded vowel spaces will show less vowel-to-vowel coarticulation than languages with less crowded vowel spaces. Manuel based her hypothesis on three assumptions:

1. There are output constraints on how a given phone can be articulated.
2. Output constraints are affected by language-particular systems of phonetic contrast.
3. Coarticulation is limited in a way that respects those output constraints.

The first two assumptions are inherent to the DIVA Speech Recognition System, since this system recognizes acceptable sounds in a language-specific manner. This leads to the learning of speech targets that embody the output constraints; i.e., the range along each orosensory dimension of a convex region target encodes the acceptable amount of variability for that sound in the infant's native language. The interpretation of coarticulation outlined in this section effectively implements the third assumption, that coarticulation is limited in a way that respects the output constraints. Manuel (1990, p. 1286) suggests that "speakers generally limit coarticulation such that it does not destroy the distinctive attributes of gestures." In contrast to coproduction models that have no clear means to guarantee that competing influences on the articulators do not destroy distinctive attributes, the generalized look-ahead model insures that this does not happen because the model coarticulates for future phonemes only when it can while still remaining within the convex region of the current phoneme. As described above, this leads to less coarticulation when the target ranges are smaller (i.e., when the output constraints are more strict). The cross-linguistic studies of Manuel (1990) and Manuel and Krakow (1984) support this result: languages with more crowded vowel spaces (and thus smaller vowel convex region targets) showed less coarticulation than languages with less crowded vowel spaces.

Several other factors can lead to smaller convex region targets. In Section 7, a method for shrinking the convex region target to produce clearer speech when rate constraints are less stringent was outlined. Similarly, in noisy conditions or when speaking to chidren or non-native listeners, speakers tend to "overarticulate" (Lindblom and MacNeilage, 1986; Manuel, 1990), involving a slowing down of speaking rate and most likely a sharpening of the vocal tract target (Picheney, Durlach, and Braida, 1985, 1986). Lindblom (1990) proposed

that speakers use a continuum from hypoarticulation to hyperarticulation when varying between casual speech and formal speech, and De Jong, Beckman, and Edwards (1993) concluded that stressed syllables were also produced by a process of hyperarticulation.

Manuel (1990, p. 1295) points out that such examples imply "that, at some level, speakers have an awareness of the notion of 'best production' and the range of acceptable productions." The convex region target for a phoneme encodes the range of acceptable production, and the notion of "best" production is implemented in DIVA by the use of a non-specific input to the ODV cells to shrink the size of convex region targets as described in Section 7. Furthermore, the generalized look-ahead model suggests that this shrinking of convex region size should lead to less coarticulation, analogous to the studies of Manuel (1990) and Manuel and Krakow (1987). This seems to be the case in human speakers, as De Jong, Beckman, and Edwards (1993) reported that subjects showed less coarticulation when producing stressed syllables than when producing unstressed syllables.

Another interesting prediction of the model concerns the speech of young children who have not yet fully learned the acceptable ranges of variability for all phonemes. In the learning process described in Section 3, the convex region targets for a speech sound start out very small and are expanded to encompass the entire range of variability allowed for the speech sound. Children who are still learning the full range of variability would possess smaller convex region targets than adults, and consequently the generalized look-ahead model predicts less coarticulation in children. Several pieces of data suggest that this is indeed the case; e.g., younger children tend to use less anticipatory nasal coarticulation (Thompson and Hixon, 1979), less anticipatory movement of the tongue body during vowels (Kent, 1983), and less anticipatory coarticulation of place during velar stops (Sereno and Lieberman, 1987).

Finally, the treatment of coarticulation in the current model can be compared with that of the neural network models of Jordan (1986, 1990). Jordan (1986) defines a recurrent back-propagation model that can be used to learn a time course of distinctive features corresponding to a phoneme string. Through the use of "don't care" terms in the teaching vectors for the model, anticipatory coarticulation is shown to arise when the model later performs its learned phoneme string, in a manner similar to the look-ahead theory. Jordan (1990) describes a second neural network model that addresses the issue of coarticulation, this time from the viewpoint of motor learning as a constrained optimization problem.

This work describes how articulatory space smoothness constraints implemented during learning can lead to anticipatory coarticulation, even in cases where task space distinctiveness constraints are used to maximize the distinctiveness of the perceptual results of different tasks (e.g., phonemic distinctiveness). This modeling work gives insight into why and how coarticulatory behavior arises in systems that learn to minimize effort by maximizing movement smoothness.

Although the mechanistic differences between the work of Jordan (1986, 1990) and the current work are too numerous to discuss here, the most important difference between the two modeling programs is a difference in scope. Whereas the goals of Jordan (1986, 1990) were the elucidation of general concepts of motor learning and performance, the goals of the current work are to provide a detailed account of a single motor behavior, speech production. Therefore, the current work addresses not only anticipatory coarticulation but also motor equivalence, velocity/distance relationships, speaking rate effects, and carryover coarticulation within a single modeling framework. Research efforts to synthesize key aspects of the two approaches may lead to a more complete description of coarticulation in speech production, e.g. through the incorporation of smoothness constraints as studied by Jordan (1990) into the DIVA learning and performance processes.

10. Concluding Remarks

As Levelt (1989; p. 452) insightfully remarks about the speech production literature, "There is no lack of theories, but there is a great need of convergence." This article has shown that study of the process by which infants learn to control their speech articulators leads to many important theoretical contributions to the ongoing process of understanding speech production. This was possible because speech acquisition was studied within the framework of a computational model of speech production, rather than in isolation. Theoretical convergence is not gained by addressing problems such as speech sound acquisition, motor equivalence, coarticulation, speaking rate effects, and variability of articulator movements separately; only by studying these phenomena within a common modeling framework can maximal convergence be attained. Because the dynamics of such a model are necessarily complex and its properties are typically difficult to clearly visualize, objective verification of the model's properties must also be possible. To meet these requirements, the current model was formulated as an adaptive neural network whose speech

production properties were verified through computer simulation. This model brings together contributions from many researchers, including the use of an action-perception or babbling cycle to tune model parameters (Bullock et al., 1993; Gaudiano and Grossberg, 1991), the use of coordinative structures (Easton, 1972; Fowler, 1980; Saltzman and Munhall, 1989), the use of orosensory information for target specification (Lindblom et al., 1979; Perkell, 1980), the incorporation of constriction locations and degrees in this target specification (Saltzman and Munhall, 1989), the use of target ranges rather than positions (Keating, 1990; Manuel, 1987, 1990), the use of a continuum from hyperarticulation to hypoarticulation (Lindblom, 1990), the use of a look-ahead process for anticipatory coarticulation (Henke, 1966; Kozhevnikov and Chistovich, 1965), the incorporation of a multiplicative gating signal for volitional speed control (Bullock and Grossberg, 1988), and the use of a direction-to-velocity mapping to gain motor equivalence capabilities (Bullock et al, 1993; Guenther, 1992).

Investigating how an infant can learn a mapping from desired movement trajectories formulated in an orosensory coordinate frame into the motor coordinate frame of articulator movements led to a simplified solution to the inverse kinematics problem for a redundant system. This solution provides a natural explanation for the formation of coordinative structures, and simulations verified motor equivalent properties seen in human speech such as automatic compensation for articulator constraints and perturbations. Data on the direct relationship between movement distance and peak movement velocity were also explained as a result of this mapping, including differences in the slope of this relationship for different sound classes (i.e., fricatives vs. stops) and for different speaking rates.

Addressing the question of how the nervous system learns which orosensory information is important for a particular speech sound resulted in a new convex region theory of the targets of speech. This theory generalizes and extends the window theory of coarticulation posited by Keating (1990), while addressing shortcomings pointed out by Fowler (1990) and Keating herself, who offered no procedure for constructing articulator paths through window targets. Convex region targets were shown to provide an intuitive explanation for data on variability in speech production, and simulations verified the model's ability to explain these data.

The implications of the convex region theory on several long-studied speech production phenomena were then investigated. It was first shown that this theory provides an insight-

ful and parsimonious explanation for a collection of speaking rate effects not previously treated by a single model. A simple non-specific input to ODV cells can be used to shrink the size of convex region targets for clearer speech at slower speaking rates, in accordance with data on vowel reduction and the speed-accuracy trade-off described by Fitts' Law. Even though the same process is used for producing vowels and consonants, differential effects of increased speaking rates on the two sound types result, as seen in human speech: consonant movement velocities increase with increased speaking rate, but vowel movements increase by a smaller amount or even *decrease* with increased rate. Despite the differential effects on movement velocities, it was shown that the ratio of maximum velocity to movement distance increases by about the same amount for the two sound types, again as seen in human speaking data. Furthermore, cross-speaker differences in strategies for increasing speaking rate are captured by variation of a single parameter.

Next, data on carryover coarticulation were addressed. The convex region framework allowed several different carryover coarticulation phenomena to be classified, and simulation results verified these phenomena in the model's productions. In contrast to the view of carryover coarticulation as the result of mechano-inertial effects, carryover coarticulation in DIVA is "planned" in the sense that it results from explicit movement commands. This planning does not require advance knowledge of later segments, but instead arises from the interaction between the configuration of the vocal tract at the start of a segment and the convex region target for the segment. This explanation of carryover coarticulation accounts for the seemingly incongruous observations that carryover coarticulation can occur with knowledge only of the next segment to be produced, yet carryover effects can extend for several segments.

Finally, anticipatory coarticulation was studied within the framework of convex region targets. It was shown that current models of coarticulation assume a target type that is a special case of the convex region target which underspecifies the shape of the vocal tract. Next, the look-ahead model of coarticulation was generalized to allow for convex region targets. This generalized look-ahead approach was implemented in DIVA, and anticipatory coarticulation was verified in model simulations. Because this generalized look-ahead approach posits that the amount of coarticulation is limited by the size of the convex region targets, it accounts for experimental results showing decreased coarticulation in cases where smaller targets are necessitated, including speech in languages with more

crowded vowel spaces, hyperarticulated speech for clarity or stress, and speech of small children who may have not yet learned the full range of variation allowed for some phonemes.

In closing, it should be noted that the model as posited here does not address many important issues concerning the control of timing in speech production (e.g., Fowler, 1980). For example, some phonemic segments, such as dipthongs and glides, are defined by the motions and rates of motions of the articulators, rather than by static configurations of the vocal tract. This suggests generalization of the convex region targets to be spatio-temporal rather than simply spatial; i.e., each segment's target is a convex region whose shape can vary with time. Ongoing research includes an investigation of these timing issues as well as the incorporation of true acoustic information into the action-perception cycle.

References

Abbs, J. H. (1973). The influence of the gamma motor system on jaw movements during speech: A theoretical framework and some preliminary observations. Journal of Speech and Hearing Research, 16, 175-200.

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

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.

Abbs, J. H., & Netsell, R. (1973). Coordination of the jaw and lower lip during speech production. Paper presented at the American Speech and Hearing Association Convention, Detroit.

Adams, S. G., Weismer, G., & Kent, R. D. (1993). Speaking rate and speech movement velocity profiles. Journal of Speech and Hearing Research, 36, 41-54.

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.

Baum, S. R., & Waldstein, R. S. (1991). Perseveratory coarticulation in the speech of profoundly hearing-impaired and normally hearing children. Journal of Speech and Hearing Research, 34, 1286-1292.

Beggs, W. D. A., & Howarth, C. I. (1972). The movement of the hand towards a target. Quarterly Journal of Experimental Psychology, 24, 448-453.

Bell-Berti, F. (1980). Velopharyngeal function: A spatio-temporal model. In N. Lass (Ed.), Speech and language: Advances in basic research and practice (pp. 291-316). New York: Academic Press.

Bell-Berti, F., & Harris, K. S. (1979). Anticipatory coarticulation: Some implications from a study of lip rounding. Journal of the Acoustical Society of America, 65, 1268-1270.

Benguerel, A. P., & Cowan, H. A. (1974). Coarticulation of upper lip protrusion in French. Phonetica, 30, 41-55.

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

Boyce, S. E., Krakow, R. A., Bell-Berti, F., & Gelfer, C. E. (1990). Converging sources of evidence for dissecting articulatory movements into core gestures. Journal of Phonetics, 18, 173-188.

Boysson-Bardies, B. de, Halle, P., Sagart, L., & Durand, C. (1989). A crosslinguistic investigation of vowel formants in babbling. Journal of Child Language, 16, 1-17.

Boysson-Bardies, B. de, Sagart, L., & Durand, C. (1984). Discernible differences in the babbling of infants according to target age. Journal of Child Language, 11, 1-15.

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

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.

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.

Caminiti, R., Johnson, P. B., & Urbano, A. (1990). Making arm movements within different parts of space: Dynamic aspects in the primate motor cortex. Journal of Neuroscience, 10, 2039-2058.

Chomsky, N., & Halle, M. (1968). The sound pattern of English. New York: Harper and Row.

Cohen, M. A., Grossberg, S., & Stork, D. G. (1988). Speech perception and production by a self-organizing neural network. In Y. C. Lee (Ed.), Evolution, learning, cognition, and advanced architectures. Hong Kong: World Scientific Publishers.

Craig, J. J. (1986). Introduction to Robotics: Mechanics and Control. Reading, MA: Addison-Wesley.

Daniloff, R., & Hammarberg, R. E. (1973). On defining coarticulation. Journal of Phonetics, 1, 239-248.

Daniloff, R., & Moll, K. (1968). Coarticulation of lip rounding. Journal of Speech and Hearing Research, 11, 707-721.

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

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

De Nil, L. F., & Abbs, J. H. (1991). Influence of speaking rate on the upper lip, lower lip, and jaw peak velocity sequencing during bilabial closing movements. Journal of the Acoustical Society of America, 89, 845-849.

Easton, T. A. (1972). On the normal use of reflexes. American Scientist, 60, 591-599.

Fitts, P. M. (1954). The information capacity of the human motor system in controlling the amplitude of movement. Journal of Experimental Psychology, 47, 381-391.

Fitts, P. M., & Peterson, J. R. (1964). Information capacity of discrete motor responses. Journal of Experimental Psychology, 67, 103-112.

Flege, J. E. (1988a). Effects of speaking rate on tongue position and velocity of movement in vowel production. Journal of the Acoustical Society of America, 84, 901-916.

Flege, J. E. (1988b). Anticipatory and carry-over nasal coarticulation in the speech of children and adults. Journal of Speech and Hearing Research, 31, 525-536.

Flege, J. E. (1991). Age of learning affects the authenticity of voice onset time (VOT) in stop consonants produced in a second language. Journal of the Acoustical Society of America, 89, 395-411.

Flege, J. E. (1993). Production and perception of a novel, second-language phonetic contrast. Journal of the Acoustical Society of America, 93, 1589-1608.

Flege, J. E., & Eefting, W. (1988). Imitation of a VOT continuum by native speakers of English and Spanish: Evidence for phonetic category formation. Journal of the Acoustical Society of America, 83, 729-740.

Folkins, J. W., & Abbs, J. H. (1975). Lip and jaw motor control during speech: Responses to resistive loading of the jaw. Journal of Speech and Hearing Research, 18, 207-220.

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

Fowler, C. A. (1990). Some regularities of speech are not consequences of formal rules: Comments on Keating's paper. In J. Kingston & M. E. Beckman (Eds.), Papers in laboratory phonology I: Between the grammar and physics of speech (pp. 476-487). Cambridge: Cambridge Univ. Press.

Fowler, C. A., & Saltzman, E. (1993). Coordination and coarticulation in speech production. Language and Speech, 36, 171-195.

Freund, H. J., & Büdingen, H. J. (1978). The relationship between speed and amplitude of the fastest voluntary contractions of human arm muscles. Experimental Brain Research, 31, 1-12.

Gaudiano, P., & Grossberg, S. (1991). Vector associative maps: Unsupervised real-time error-based learning and control of movement trajectories. Neural Networks, 4, 147-183.

Gay, T. (1977). Articulatory movements in VCV sequences. Journal of the Acoustical Society of America, 62, 183-193.

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

Georgopoulos, A. P., Kalaska, J. F., Caminiti, R., & Massey, J. T. (1982). On the relations between the direction of two-dimensional arm movements and cell discharge in primate motor cortex. Journal of Neuroscience, 2, 1527-1537.

Georgopoulos, A. P., Kalaska, J. F., & Massey, J. T. (1981). Spatial trajectories and reaction times of aimed movements: Effects of practice, uncertainty, and change in target location. Journal of Neurophysiology, 46, 725-743.

Ghez, C., & Vicario, D. (1978). The control of rapid limb movement in the cat, II: Scaling of isometric force adjustments. Experimental Brain Research, 33, 191-202.

Gopal, H. S. (1990). Effects of speaking rate on the behavior of tense and lax vowel durations. Journal of Phonetics, 18, 497-518.

Grobstein, P. (1991). Directed movement in the frog: A closer look at a central representation of spatial location. In M. A. Arbib & J. P. Ewert (Eds.), Visual structures and integrated functions (pp. 125-138). Berlin: Springer-Verlag.

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

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

Henke, W. L. (1966). Dynamic articulatory model of speech production using computer simulation. Unpublished doctoral dissertation, Massachusetts Institute of Technology, Cambridge, MA.

Hirayama, M., Vatikiotis-Bateson, E., Kawato, M., and Jordan, M. I. (1992). In J. E. Moody, S. J. Hanson, & R. P. Lippmann (Eds.), Advances in neural information processing systems 4 (pp. 191-198). San Mateo, CA: Morgan Kaufmann Publishers.

Houde, R. A. (1967). A study of tongue body motion during selected speech sounds. Unpublished doctoral dissertation, University of Michigan, Ann Arbor, MI.

Howarth, C. I., & Beggs, W. D. A. (1971). The relationship between speed and accuracy of movement aimed at a target. Acta Psychologica, 35, 207-218.

Jagacinski, R. J., & Monk, D. L. (1985). Fitts' Law in two dimensions with hand and head movements. Journal of Motor Behavior, 17, 77-95.

Jakobson, R., & Halle, M. (1956). Fundamentals of language. The Hague: Mouton.

Jordan, M. I. (1986). Serial order: A parallel distributed processing approach (Tech Report ICS 8604). San Diego: University of California San Diego.

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.

Kaplan, E., & Kaplan, G. (1971). The prelinguistic child. In J. Eliot (Ed.), Human development and cognitive processes (pp. 358-381). New York: Holt, Rinehart, and Winston.

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

Kelso, J. A. S., Tuller, B., Vatikiotis-Bateson, E., & Fowler, C. A. (1984). Functionally specific articulatory cooperation following jaw perturbations during speech: Evidence for coordinative structures. Journal of Experimental Psychology: Human Perception and Performance, 10, 812-832.

Kent, R. D. (1983). The segmental organization of speech. In P. F. MacNeilage (Ed.), The production of speech (pp. 57-89). New York: Springer-Verlag.

Kent, R. D., Carney, P., & Severeid, L. (1974). Velar movement and timing: Evaluation of a model for binary control. Journal of Speech and Hearing Research, 17, 470-488.

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

Kent, R. D., & Moll, K. L. (1972a). Cinefluorographic analyses of selected lingual consonants. Journal of Speech and Hearing Research, 15, 453-473.

Kent, R. D., & Moll, K. L. (1972b). Tongue body articulation during vowel and diphthong gestures. Folia Phoniatrica, 24, 278-300.

Knight, A. A., & Dagnall, P. R. (1967). Precision in movements. Ergonomics, 10, 327-330.

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

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

Kuehn, D. P. (1973). A cinefluorographic investigation of articulator velocities. Unpublished doctoral dissertation, University of Iowa, Iowa City, IA.

Kuehn, D. P., & Moll, K. L. (1976). A cineradiographic study of VC and CV articulatory velocities. Journal of Phonetics, 4, 303-320.

Kuhl, P. K. (1979). Speech perception in early infancy: Perceptual constancy for spectrally dissimilar vowel categories. Journal of the Acoustical Society of America, 66, 1668-1679.

Kuhl, P. K., Williams, K. A., Lacerda, F., Stevens, K. N., & Lindblom, B. (1992). Linguistic experience alters phonetic perception in infants by 6 months of age. Science, 255, 606-608.

Levelt, W. J. M. (1989). Speaking: From intention to articulation. Cambridge, MA: MIT Press.

Lindblom, B. (1963). Spectrographic study of vowel reduction. Journal of the Acoustical Society of America, 35, 1773-1781.

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

Lindblom, B. (1990). Explaining phonetic variation: A sketch of the H&H theory. In H. J. Hardcastle & A. Marchal (Eds.), Speech production and speech modeling (pp. 403-440). Dordrecht, Holland: Kluwer.

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.

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

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

Lynch, M. P., & Oller, D. K. (1989). Development of speech-like vocalizations in a child with congenital absence of cochleas: The case of total deafness. Applied Psycholinguistics, 10, 315-333.

MacNeilage, P. F. (1970). Motor control of serial ordering in speech. Psychological Review, 77, 182-196.

MacNeilage, P. F., & Davis, B. (1990). Acquisition of speech production: Frames, then content. In M. Jeannerod (Ed.), Attention and performance XIII: Motor representation and control (pp. 453-476). Hillsdale, NJ: Erlbaum.

MacNeilage, P. F., & Ladefoged, P. (1976). The production of speech and language. In E. C. Carterette & M. P. Friedman (Eds.), Handbook of perception, volume VII: Language and speech (pp. 76-120). New York: Academic Press.

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.

Manuel, S. Y. (1987). Acoustic and perceptual consequences of vowel-to-vowel coarticulation in three Bantu languages. Unpublished doctoral dissertation, Yale University, New Haven, CT.

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

Manuel, S. Y., & Krakow, R. A. (1984). Universal and language particular aspects of vowel-to-vowel coarticulation (Haskins Laboratory Status Report on Speech Research SR-77/78: 69-78). New Haven, CT: Haskins Laboratory.

Miyawaki, K., Strange, W., Verbrugge, R., Liberman, A. M., Jenkins, J. J., & Fujimura, O. (1975). An effect of linguistic experience: The discrimination of [r] and [l] by native speakers of Japanese and English. Perception and Psychophysics, 18, 331-340.

Munhall, K. G., Ostry, D. J., & Flanagan, J. R. (1991). Coordinate spaces in speech planning. Journal of Phonetics, 19, 293-307.

Öhman, S. E. G. (1966). Coarticulation in VCV utterances: Spectrographic measurements. Journal of the Acoustical Society of America, 39, 151-68.

Öhman, S. E. G. (1967). Numerical model of coarticulation. Journal of the Acoustical Society of America, 41, 310-320.

Oller, D. K. (1980). The emergence of the sounds of speech in infancy. In G. H. Yeni-Komshian, J. F. Kavanagh, & C. A. Ferguson (Eds.), Child phonology, volume 1: Production (pp. 93-112). New York: Academic Press.

Oller, D. K., & Eilers, R. E. (1988). The role of audition in infant babbling. Child Development, 59, 441-449.

Ostry, D. J., & Munhall, K. G. (1985). Control of rate and duration of speech movements. Journal of the Acoustical Society of America, 77, 640-648.

Penfield, W., & Rasmussen, T. (1950). The cerebral cortex of man: A clinical study of localization and function. New York: MacMillan.

Perkell, J. S. (1969). Physiology of speech production: Results and implications of a quantitative cineradiographic study. Research Monograph No. 53. Cambridge, MA: MIT Press.

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.

Perkell, J. S., & Nelson, W. L. (1982). Articulatory targets in speech motor control: A study of vowel production. In S. Grillner, A. Persson, B. Lindblom, & J. Lubker (Eds.), Speech Motor Control. New York: Pergamon.

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

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

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

Recasens, D. (1987). An acoustic analysis of V-to-C and V-to-V coarticulatory effects in Catalan and Spanish VCV sequences. Journal of Phonetics, 15, 299-312.

Recasens, D. (1989). Long range coarticulation effects for tongue dorsum contact in VCVCV sequences. Speech Communication, 8, 293-307.

Rootes, T. P., and MacNeilage, P. F. (1967). Some speech perception and production tests of a patient with impairment in somesthetic perception and motor function. In J. F. Bosma (Ed.), Symposium on Oral Sensation and Perception (pp. 310-317). Springfield, IL: Thomas.

Sachs, J. (1976). The development of speech. In E. C. Carterette & M. P. Friedman (Eds.), Handbook of perception, volume VII: Languange and speech (pp. 145-172). New York: Academic Press.

Sakata, H., Shibutani, H., & Kawano, K. (1980). Spatial properties of visual fixation neurons in posterior parietal association cortex of the monkey. Journal of Neurophysiology, 43, 1654-1672.

Saltzman, E. L., & Kelso, J. A. S. (1987). Skilled actions: A task-dynamic approach. Psychological Review, 94, 84-106.

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

Schmidt, R. A. (1982). Motor control and learning: A behavioral emphasis. Champaign, IL: Human Kinetics Publishers.

Sereno, J. A., & Lieberman, P. (1987). Developmental aspects of lingual coarticulation. Journal of Phonetics, 15, 247-257.

Stark, R. E. (1980). Stages of speech development in the first year of life. In G. H. Yeni-Komshian, J. F. Kavanagh, & C. A. Ferguson (Eds.), Child phonology, volume 1: Production (pp. 73-92). New York: Academic Press.

Sussman, H. M., & Smith, J. U. (1971). Jaw movements under delayed auditory feedback. Journal of the Acoustical Society of America, 50, 685-691.

Thompson, A. E., & Hixon, T. J. (1979). Nasal air flow during normal speech production. Cleft Palate Journal, 16, 412-420.

Whalen, D. H. (1990). Coarticulation is largely planned. Journal of Phonetics, 18, 3-35.

Wickelgren, W. A. (1969). Context sensitive coding, associative memory, and serial order in (speech) behavior. Psychological Review, 76, 1-15.

Wood, S. A. J. (1991). X-ray data on the temporal coordination of speech gestures. Journal of Phonetics, 19, 281-292.

Woodworth, R. S. (1899). The accuracy of voluntary movement. Psychological Review, 3, 1-114.

Zelaznik, H. N., Schmidt, R. A., & Gielen, S. C. A. M. (1986). Kinematic properties of rapid aimed hand movements. Journal of Motor Behavior, 18, 353-372.

Zlatin, M. A., & Koenigsknecht, R. A. (1976). Development of the voicing contrast: A comparison of voice onset time in stop perception and production. Journal of Speech and Hearing Research, 19, 93-111.