BOSTON UNIVERSITY

GRADUATE SCHOOL OF ARTS AND SCIENCES

Dissertation

INVESTIGATIONS OF FORMANT AND WAVELET REPRESENTATIONS FOR SPEECH MOVEMENT PLANNING

by

CHARLES DAVID JOHNSON

B.S., University of Illinois, 1979
M.S., Santa Clara University, 1983

Submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

1998

Approved by

First Reader

Frank H. Guenther, PhD
Assistant Professor of Cognitive and Neural Systems

Second Reader

Michael A. Cohen, PhD
Associate Professor of Cognitive and Neural Systems

Third Reader

Stephen Grossberg, PhD
Wang Professor of Cognitive and Neural Systems

Acknowledgments

I would especially like to thank Frank Guenther, my advisor, who invited me to collaborate with him on speech production modeling, and who supported and encouraged me every step of the way.

I would also like to thank the faculty, staff, and students of the CNS department for everything they've done to make this work possible, especially Stephen Grossberg and Gail Carpenter, for their vision and commitment to give us this wonderful interdisciplinary department. Grateful acknowledgement is also given for support by the Office of Naval Research and the National Institutes of Health[1].

My dream to study in the Department of Cognitive and Neural Systems grew out of a trip to the Wang Institute at Lowell, Massachusetts, in the spring of 1990, to attend a weeklong neural network conference sponsored by the CNS department. That conference was instrumental in my decision to invest in the study of neural networks, and convinced me that BU was the place to do it. Thanks to Ennio Mingolla who took time out of his busy week to discuss the graduate program with me and to encourage me to apply. The prospect of returning to full-time graduate study after 14 successful years of engineering work in the computer industry was terrifying. I gratefully acknowledge the encouragement of my closest friends, Staci Robbins Silber, Bill Smith, Neil Schulte, and especially Lynn Sharp, who gave me the strength to pursue this dream. I would also like to thank my mother, Patricia Ann Johnson, and my father and his wife, Melvin Edward and Elizabeth Lowe Johnson, for their many positive contributions to my life and especially for their generous support over the past two years. Also, thanks to my brothers, Steve, Doug, and Dan, who through the

[1]AASERT (ONR N00014-94-1-0940), NIH 1-R29-DC02952-01, ONR N00014-95-1-0657, ONR N00014-94-1-0597, and MURI (ONR N00014-95-1-0409).

years have continued to push me in so many positive ways.

Some people will wonder how I came to apply wavelets to the problem of speech production modeling. During a collaborative research effort with Stephen Grossberg on the visual perception of curved surfaces in depth, I conducted a literature review of multiscale methods in image processing and biological vision, and began to learn the mathematical intricacies of the wavelet formalism. Wavelet methods were finding wide application in image processing and, to a lesser extent, in biological modeling of visual perception, and I began to apply these methods to visual perception modeling. In October, 1996, Shihab Shamma gave a CNS colloquium on the wavelet representation of sound spectra in auditory cortex, which inspired me to find a way to use this representation in a speech production system. I gratefully acknowledge Shihab Shamma's contributions to my research.

Thanks to Marcos Campos for our many long and fruitful discussions on all aspects of wavelet and multiscale methods, and to Lars Lidén who tirelessly acted as a sounding board for many of the ideas presented herein. Thanks to Michael Cohen for his thoughtful criticism of every paragraph of this dissertation. Also, thanks to Diana Meyers for being a supportive friend and for helping with the final submission of the dissertation to the graduate school after I had taken a job in Austin, Texas. Also, thanks to Scott Oddo, Michelle Hampson, Siegfried Martens, and Greg Gancarz, for their friendship through the best and worst of times!

Finally, a warmest possible expression of gratitude to Lynn Sharp for inspiring and encouraging me to return to graduate school and for keeping me afloat during the hardest first two years. Without her active involvement, this dream would have remained nebulous and distant.

to my family

INVESTIGATIONS OF
FORMANT AND WAVELET REPRESENTATIONS
FOR SPEECH MOVEMENT PLANNING

(Order No.)

CHARLES DAVID JOHNSON

Boston University Graduate School of Arts and Sciences, 1998

Major Professor: Frank H. Guenther, Assistant Professor of Cognitive and Neural Systems

Abstract

Models of speech production have historically assumed that phoneme targets are represented as points in a planning space of vocal tract constrictions. However, recent evidence suggests that speakers may utilize an acoustic-like space in the planning of articulator movements for vowel and consonant production. This thesis demonstrates the feasibility of using acoustic planning spaces for the production of vowel sounds in DIVA, a computational neural model of motor-equivalent vowel production. Two acoustic planning spaces are studied, a planning space based on formant frequencies and a planning space based on a wavelet decomposition of the speech spectrum. The successful implementation of the formant planning space proves that articulatory control based on formants is practical and does not suffer from inherent nonlinearities in the inverse kinematic map between desired formant movements and corresponding movements of the vocal articulators. However, formant-based models leave a number of issues unresolved, and several researchers have suggested that the gross shape of the vowel spectrum may correlate more closely with vowel perception data. Recent physiological studies suggest that the peripheral auditory system computes the log magnitude spectrum of a steady vowel, and that the representation in primary audi-

tory cortex is similar to a wavelet decomposition of this spectrum. Based on these psychophysical and physiological results, the thesis proposes a model of vowel production based on a wavelet expansion of the log magnitude spectrum of the target vowel. The model employs an orthonormal set of wavelet basis functions which spans the space of possible vowel spectra. The wavelet auditory planning space dimensions correspond to the coefficients in the wavelet expansion of the spectrum, and vowel targets are assumed to be connected regions in this space. In addition to support from the physiological literature, this model has a number of advantages over formant-based vowel production models, including simpler computation of the spectral parameters and better approximations of gross spectral shape. Also, the wavelet auditory planning space is used to explain the spectral center of gravity effect, in which formant clusters are averaged into a single formant peak during vowel perception.

Contents

List of Tables

List of Figures

Chapter 1

Introduction: Reference Frames for Speech Production

A number of different spaces, or reference frames are involved in the production of speech. These include the phonetic frame which codes the phonemes to be produced, the acoustic frame which codes the acoustic attributes of a phoneme, the orosensory frame which codes the tactile and proprioceptive feedback from the oral cavity, the articulatory frame which codes the articulatory movements necessary to produce a phoneme, and the planning frame in which the utterance is planned (Guenther, 1995b; Guenther, Hampson, & Johnson, 1998). It is useful to think of speech production as the process of forming a trajectory in the planning space so that the trajectory passes through a sequence of targets, each corresponding to a different phoneme in a phoneme string. Movements along this trajectory can then be mapped into movements of articulators that carry out the planned trajectory (Guenther, 1995a). A computational model of speech production must describe these coordinate frames and their interactions in sufficient detail to permit generation of a speech utterance.

Current research in speech production is being fueled by two important debates in the speech literature. One long-standing debate has been concerned with the nature of targets in the planning space, and has sought to answer the question: Is the speech production planning target constriction-like or acoustic-like? Several recent models have used reference frames that correspond to the locations and degrees of certain

key constrictions in the vocal tract. The task-dynamic model (Saltzman & Munhall, 1989) and the original formulation of the DIVA model (Guenther, 1994a, 1995b) use constriction-based planning spaces and are capable of motor-equivalent speech production.

However, recent evidence suggests that speakers utilize a more acoustic-like space for planning vowel movements. The evidence that speech production is planned in acoustic-like space comes largely from experiments on compensatory behavior during speech production (Perkell, Matthies, Svirsky, & Jordan, 1993; Savariaux, Perrier, & Orliaguet, 1995; de Jong, 1997). For example, Perkell et al. (1993) studied production of the vowel /OO/ and hypothesized that "[t]he objective of articulatory movements is an acoustic goal" (page 2948), rather than a goal more closely related to the articulators such as a constriction goal, based on experimental results indicating that speakers use trade-offs in constriction parameters (lip rounding and tongue-body raising) to reach an acoustic goal such as a target value of the second formant frequency (F2). Analogous results have recently been observed for consonant production (Perkell, Matthies, & Svirsky, 1994). These results suggest that speakers are not planning movements to constriction targets, but instead are planning movements toward acoustic targets. This in turn suggests that speech movements are planned in a more acoustic-like reference frame. This makes sense since the true goal of the speech production system is the creation of an acoustic signal that can be properly interpreted by listeners, not the production of specific constrictions in the vocal tract.

Another debate has, so far, been primarily limited to the speech perception literature, and is concerned with the nature of the acoustic correlates of vowel and consonant perception. Traditional theories of vowel perception hold that vowels are

distinguished by the locations of spectral peaks or *formants* (Peterson & Barney, 1952; Ladefoged & Broadbent, 1957; Lindblom & Studdert-Kennedy, 1967; Miller, 1989). Formants have played a major role in theories of vowel perception and production. A formant is a relative maximum in the acoustic spectrum of the vowel, and is typically characterized by its center frequency and amplitude (or bandwidth). It has been shown that a formant amplitude and its bandwidth are mathematically related in a simple fashion when there are no zeros in the vocal tract transfer function (Flanagan, 1957; Klatt, 1980), which is usually the case for nonnasalized vowels. In addition, it has been shown that peaks in the spectrum of vowels are perceptually more important than the between-peak amplitudes (e.g., Flanagan, 1957). Consequently, a representation of the spectrum based on formants is usually adequate for single-speaker vowel recognition and production.

However, recent psychophysical (Zahorian & Jagharghi, 1993) and physiological (Shamma, 1988) evidence suggests that gross spectral shape may be more important for distinguishing vowels. Another reason for investigating alternatives to formant-based models of speech perception and production is that extraction of formant information from continuous speech is very difficult and unreliable (Flanagan, 1956; Zahorian & Jagharghi, 1993). Many of the same issues that occur in the debate between the proponents of formants and the proponents of gross spectral shape in the speech perception literature can also be expected to manifest themselves in the speech production domain. One goal of the dissertation is to expose some of these issues, and shed some light on the role of formants and gross spectral shape in the speech production process.

This thesis is based on the DIVA model of speech production (Guenther, 1994a, 1995b) which originally utilized constriction-based planning and was not capable of

producing acoustic output. The research reported in this thesis significantly enhanced the DIVA model by utilizing acoustic planning and by incorporating the Maeda articulatory system. The resulting model is able to learn how to produce vowel targets during a babbling phase, and produces intelligible vowel tokens. In addition, the thesis proposes an auditory planning space based on the wavelet transform, and presents results of simulations with the new model. The proposed wavelet planning space is based on recent advances in the understanding of auditory physiology, and offers several advantages over existing planning space models. Special attention is paid to vowel production, but stop consonant production is also considered.

This dissertation does not investigate dynamic features of vowel and consonant production. Static spectral information contributes only part of the information for vowel recognition. Strange (1989), for example, notes that, "... in no study using natural (as opposed to synthetic) stimuli were isolated vowels identified *more* accurately than coarticulated vowels, as would be predicted if static targets were the primary source of information for vowel identity" (page 2084). However, although feature trajectories are important secondary cues for vowel perception, static cues are more important than temporal cues for monophthongal vowels in isolated CVC words. It will be seen that very good results for vowel production (and, to a lesser extent, for stop consonants) are obtained by considering only static spectral information.

This dissertation is organized as follows. Chapter 2 reviews the DIVA model of speech production, in which motor-equivalent production of phonemes is learned by the model during a babbling phase, and describes the concept of acoustic planning of speech production. Simulations with a formant planning space in DIVA are used to show that acoustic planning of vowel production is feasible.

Chapter 3 considers problems with formant-based representations of vowel spec-

tra, motivates the concept of shape-based representations, and introduces a wavelet auditory representation of vowel spectra. The wavelet auditory representation is derived from recent physiological results from primary auditory cortex. These results suggest that the auditory system represents a sound by using a wavelet decomposition of the magnitude spectrum of the sound. This wavelet auditory representation of vowel spectra is then applied to the problem of understanding the spectral center of gravity effect, in which formant peaks that are close together are averaged into a single peak for the purpose of vowel perception. Psychophysical results concerning the spectral center of gravity effect are used to constrain the design of the basis functions employed in the wavelet auditory representation, and subsequent simulations show that the spectral center of gravity effect occurs naturally if a wavelet auditory representation underlies vowel perception.

Chapter 4 applies the wavelet auditory representation to the DIVA model and demonstrates that acoustic planning of vowel production using the wavelet auditory representation of vowel spectra is possible, and avoids the problems of extracting formants from the speech signal. The model successfully produces 9 English vowels, using vocal tract configurations typically employed by English speakers, and having spectral properties typical for these vowels. In addition, the model exhibits motor equivalence in simulations using a bite block. Finally, in simulations of /d/, /g/, and /b/, the model successfully produces vocal tract configurations having constriction locations and degrees typically seen in humans.

Finally, Chapter 5 presents a list of the proposed contributions of the dissertation and suggests a number of areas for future research.

Chapter 2

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

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

The research described in this chapter extends these prior results by investigating an acoustic-like planning space consisting of the first two formants of the speech signal in place of the constriction-based planning space used by Guenther (1994a, 1995b). Furthermore, the version of the model described here produces true acoustic output,

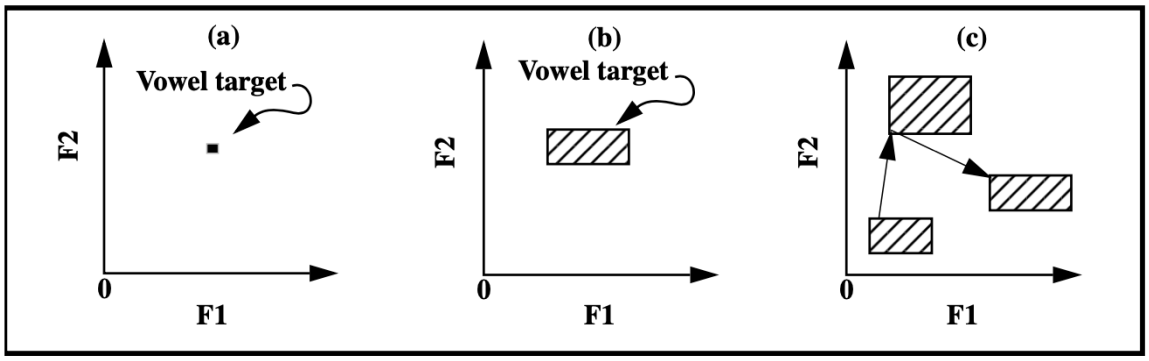


Figure 2·1: Schematization of vowel targets in F1-F2 acoustic space. (a) Earlier theories specify point targets in the planning space. (b) DIVA specifies target regions in the planning space. (c) A trajectory in planning space computed by DIVA. From a starting configuration, the minimum formant transitions required to reach a second and then a third vowel are computed by the model and mapped to the corresponding articulator directions.

which was not possible in the model of Guenther (1994a, 1995b) which used an overly simplistic articulatory structure.

The main theoretical contribution of this chapter is an existence proof that acoustic space planning (using formants), in conjunction with the Maeda articulatory system, is capable of producing a wide range of vowel sounds in the absence of any explicit specification of vocal tract shape. Because of the nonlinear relationship between vocal tract shape and the corresponding acoustic parameters of speech, there may exist vocal tract configurations at which the inverse kinematic map (described in Section 2.3.1) is zero. At such singularities, movement toward the target would stop. The research presented in this chapter demonstrates that the nonlinear inverse kinematic mapping can be learned and that local minima (e.g., zeros in the inverse kinematic mapping), if they exist, do not interfere with production of the vowel targets. The model produces vowels with vocal tract shapes similar to actual speakers.

2.1 Overview of the Model

DIVA is a model of self-organizing motor-equivalent speech production. The version of DIVA described here learns to produce ten English vowels, each characterized by a unique rectangular region in formant (F1, F2) space. Figure 2·1 illustrates the formant planning space. In Figure 2·1(a) a point target is shown. DIVA uses region vowel targets, as illustrated in Figure 2·1(b). Vowel production planning consists of computing a trajectory through the planning space from the current point in the planning space through the vowel target regions. This is schematized in Figure 2·1(c) where three vowel targets are shown, and the minimum formant transitions are computed to reach subsequent targets. In the sections that follow, the term planning space is used interchangeably with *acoustic space*, although it should be made clear that, in DIVA, the planning space is an *acoustic-like* space whose dimensions are the first two formants F1 and F2.

The model (see Figure 2·2) has two phases or modes of operation: babbling and performance. Performance of a phoneme string can be visualized as follows. The Speech Sound Map cell corresponding to the first phoneme in the string is activated. This cell's activity propagates through the weights projecting to the Planning Direction Vector, effectively "reading out" the phoneme's learned target. The Planning Direction Vector represents the difference between this target and the current state of the vocal tract; in other words, the Planning Direction Vector codes the desired movement direction in the planning space. This is then mapped into an appropriate set of articulator directions through the learned mapping from the Planning Direction Vector to the Articulator Direction Vector. As the articulators move, the corresponding acoustic state, registered through tactile and proprioceptive feedback at the Planning Position Vector stage, moves closer and closer to the acoustic target for the speech

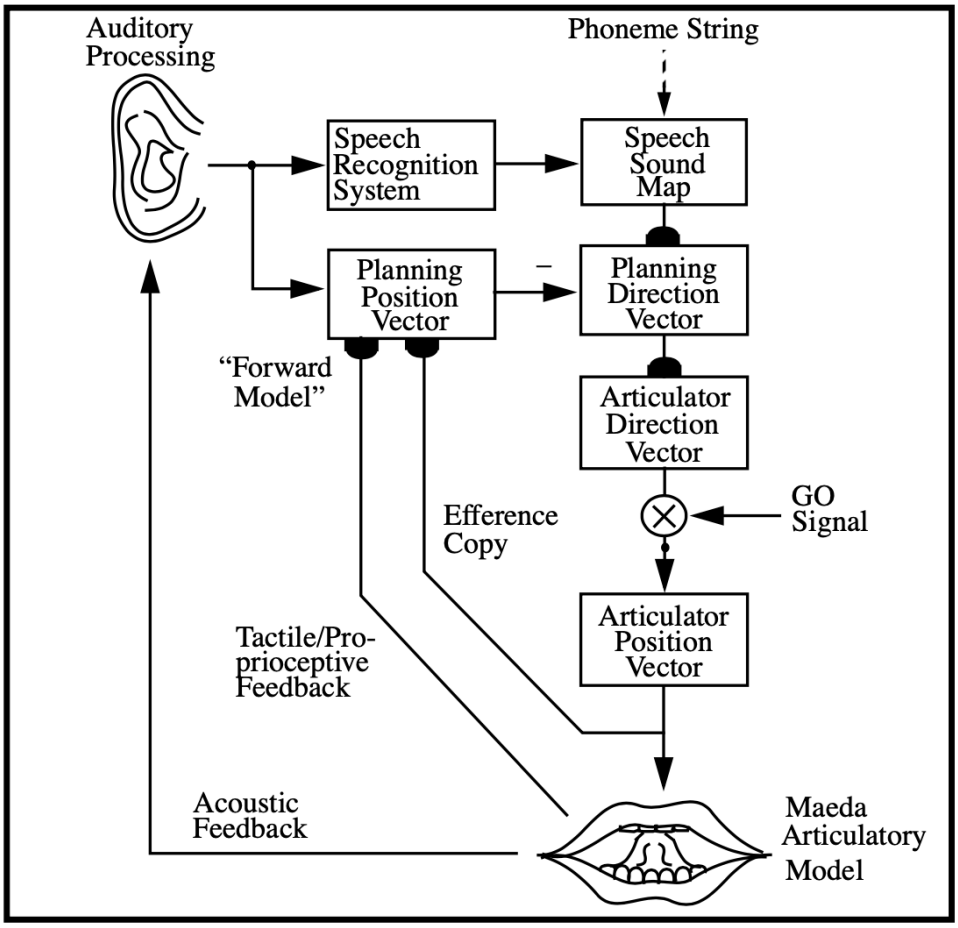


Figure 2·2: Overview of the DIVA model. During the performance phase, a desired vowel is mapped (by the phoneme-to-acoustic mapping) to its target region in the planning space. A difference vector is then computed between the position in the formant planning space (corresponding to the current vocal tract configuration) and the nearest point within the formant space target region for the desired vowel. This difference vector is then mapped (by the acoustic-to-articulatory mapping) to a direction vector in articulatory space. The Articulator Direction Vector is then integrated to obtain the Articulator Position Vector, which is analyzed to produce acoustic output. Learned mappings are indicated by filled semicircles.

sound. This causes the Planning Direction Vector activity to become smaller and smaller, leading to a slowing and stopping of articulator movements as the target is reached. These processes are carried out automatically by the temporal dynamics of the neural network. The time course of activity of the Planning Direction Vector cells can be thought of as the planned trajectory in acoustic coordinates. When Planning Direction Vector activity is sufficiently close to zero (i.e., when the sound has been successfully produced), the Speech Sound Map cell corresponding to the next phoneme in the phoneme string is activated, and the process repeats. The result is a time course of articulator positions that can be displayed as a real-time animation sequence on a computer monitor.

The babbling phase consists of two subphases. During the first subphase, the acoustic-to-articulatory mapping is learned. After the first babbling subphase completes, the second subphase begins, during which the phoneme-to-acoustic mapping is learned.

Babbling in the model is produced by inducing movements of the speech articulators by randomly activating the Articulator Direction Vector cells, which leads to movements of the speech articulators. Tactile and proprioceptive feedback provides information about the changing shape of the vocal tract within the planning reference frame (through the forward model), and acoustic feedback processed by the speech recognition system provides phonetic information. The combination of articulatory information (in the form of the randomly activated movement commands) and planning space information from the forward model allows tuning of the mapping between the Planning Direction Vector and the Articulator Direction Vector. The tuning process can be thought of as learning which articulator movements will move the vocal tract in a desired direction in planning space so as to allow the articulators to later

carry out planned trajectories. The combination of phonetic information from the speech recognition system and planning space information from the forward model allows tuning of the mapping between the Speech Sound Map and the Planning Direction Vector. This tuning process can be thought of as learning a target in planning space for each speech sound. When a sound is babbled, the sound's target is modified based on the position in planning space that led to production of the sound.

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

In the following sections, the model components are described, and the performance and babbling phases are elaborated.

2.2 Components of the Model

A block diagram of the DIVA model is shown in Figure 2·2. The model consists of a number of components described in the following paragraphs.

2.2.1 The Maeda Articulators and the Articulator Position Vector

Earlier implementations of the DIVA model (Guenther, 1994a, 1995b) were not designed to accurately represent the human vocal tract, but used, instead, a highly schematized representation of the articulators. They bear some relation to the human vocal apparatus, but are not suitable for the accurate acoustic modeling necessary for the production of intelligible vowel sounds. Even though the resulting shapes of the

vocal tract produced by the earlier DIVA might be qualitatively correct, the resulting acoustic parameters using these shapes will be wrong. Sounds would be produced, but those sounds wouldn't sound like human vowels.

Before describing the Maeda articulatory system used in this thesis, a brief historical review of vocal tract models is given. Three broad classes of vocal tract modeling approaches can be identified: area functions, articulatory models, and shape factor models. Examples of each approach and their relative advantages are given.

An area function gives the cross sectional area of the vocal tract as a function of distance from the glottis (or from the teeth which are fixed). From the area function, acoustic parameters such as formant frequencies and amplitudes (or bandwidths) may be calculated. In order to calculate these acoustic parameters, the vocal tract is approximated by a series of concatenated tubes with circular cross section and constant radius. At each tube junction, forward- and backward-propagating waves are generated, and the resulting acoustic output can be calculated by the methods described in Rabiner and Schafer (1978). Similarly, this system can be analyzed to produce the formant frequencies. Area functions are still commonly used to model the vocal tract, and the other types of models can always be transformed into their equivalent area function for the purposes of acoustic processing.

Early models of the vocal tract consisted only of an area function. For example, Stevens and House (1955) used a parabolic approximation of the area function to show that reasonably intelligible vowel sounds could be produced. Analyses of area functions showed that the most important parameters for vowel production are (1) place of constriction, (2) size of constriction, and (3) aspect ratio of the lips.

Using ever more accurate area functions, higher quality of speech can be obtained. However, study of the area function does not reveal very much about the underlying

speech production processes, especially from a motor control point of view. One reason for this is that, although the area function possesses many degrees of freedom (e.g., 17 or more in the Maeda model), only a few articulators are thought to control the vocal tract shape (e.g., 7 articulators in the Maeda model). Therefore, it is more advantageous to model the human speech articulators and use these to derive the acoustic output.

In an early articulatory system, described only in an short abstract, Ladefoged (1964) proposed a simple physiological model having 10 parameters. The parameters included (1) air pressure at the trachea, (2) position of the vocal cords, (3) tension of the vocal cords, (4) degree of velo-pharyngeal stricture, (5,6) coordinates of the center of the body of the tongue, (7,8) coordinates of the tip of the tongue, (9,10) protrusion and opening of the lips. The importance of this model is that it suggests how many speech articulators are necessary to control vocal tract shape.

Articulatory models allow researchers to systematically explore trading relations between vocal articulators, using a small set of simple articulators. Usually, the shapes of the articulators are fixed or are constrained to vary in a simple way. For example, Lindblom and Sundberg (1971) were the first to treat the jaw as a separate speech articulator, in an articulatory model that included a two-parameter model of tongue shape. In experiments with Swedish talkers, they found that jaw and tongue movements were traded in order to minimize deformation of the tongue body shape. These results were extended to include compensatory vowel articulation by Lindblom, Lubker, and Gay (1979).

Mermelstein (1973) proposed a simple geometric model of the vocal tract which he used for both vowels and consonants. It is one of the most widely cited quantitative models of the vocal tract. The model assumes that the tongue body conforms to a

circular arc of constant radius. The jaw is allowed to undergo only angular displacements. The tongue blade is a straight line. However, Mermelstein was able to fit a wide variety of acoustic data with this model. Rubin, Baer, and Mermelstein (1981) developed their extension to the Mermelstein (1973) model to serve as a tool for studying "linguistically and perceptually significant aspects of articulatory events" (page 321), which are difficult to study in terms of the area function.

Another articulatory model is due to Liljenkrants (1971), cited in Harshman, Ladefoged, and Goldstein (1977), who proposed a Fourier series decomposition of tongue shape. The model allows simple specification of a wide variety of tongue shapes, but it is unlikely that the brain codes vocal tract shape using the infinite duration sines and cosines implied by the Fourier series representation.

The third type of vocal tract model is the shape factor model. Shape factors are determined from a statistical analysis (e.g., principal components analysis) of many tracings of the vocal tract. In order to motivate the discussion of factor analysis, assume that many vocal tract shapes are encoded in vector form. In this analysis, a shape vector has a large number of components, each corresponding to a point in some suitably chosen coordinate system. The underlying assumption of factor analysis is that (under certain conditions) there exist a set of *shape factors* such that each vocal tract shape vector can be written as a linear combination of the shape factors.

Most recent models of vocal tract shape are derived using factor analysis. For example, Harshman et al. (1977) performed factor analysis of tongue shapes from vowel productions of five speakers of English and derived two tongue shape factors. These factors were used by the speakers to different degrees, possibly because of individual anatomy. Jackson (1988) studied cross-linguistic factors of shape to propose that such factors might be related to coordinative structures (Fowler, 1980), i.e., hypothesized

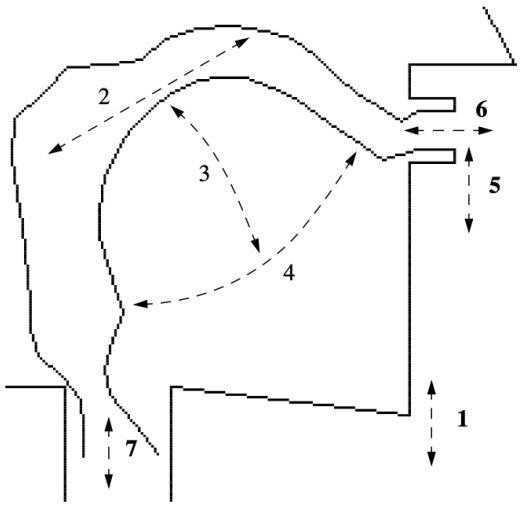


Figure 2·3: Maeda system of vocal articulators. 7 Degree-of-Freedom redundant articulatory system. (1) Jaw height. (2) Tongue-body position. (3) Tongue-body shape. (4) Tongue-tip position. (5) Lip aperture. (6) Lip protrusion. (7) Larynx height. The degrees of freedom are derived from French talkers by principal components analysis. Maeda's algorithm is used to transform an articulator configuration to the corresponding spectrum and formant values.

motor synergies of speech articulation. He suggested that some languages may favor one gesture over another and that factor analysis may be able to decide this question.

In this thesis, the DIVA model uses an articulatory system developed by Shinji Maeda (1990). Maeda (1972) suggested that shape factors derived by the method of Analysis by Synthesis may be used to capture more accurately the shapes of the vocal tract. Maeda (1990) derived an articulatory model of the vocal tract from factor analysis of cineradiographic and labiofilm data from adult French talkers. The Maeda articulatory system was chosen because it is based on analyses of vocal tract configurations obtained during the production of vowels, because the articulatory system is simple while conforming to the most important vocal tract articulators, and because a computer implementation of the articulatory system was available.

The Maeda articulatory model defines seven shape parameters or "articulators",

shown in Figure 2·3: (1) jaw height, (2) tongue-body position, (3) tongue-body shape, (4) tongue-tip position, (5) lip height (aperture), (6) lip protrusion, and (7) larynx height. Although the Maeda articulators are not biologically derived, they are similar to (and named after) the corresponding articulators in the human vocal tract. Each Maeda articulator takes on a real value in the interval [-3, 3] and may be regarded as a coefficient that weights a shape eigenvector. The sum of these weighted eigenvectors is a vector of points in the midsagittal plane and defines the outline of the vocal tract shape. The Maeda articulatory space is the set of all possible 7-tuples of Maeda articulator values, and defines a seven-dimensional, Euclidean space. Each vocal tract configuration corresponds to exactly one point in Maeda articulatory space.

To illustrate the effect of the Maeda articulators on tongue shape and position, consider Figure 2·4, which shows four of the seven shape factors: (a) jaw position, (b) tongue-dorsal position, (c) tongue-dorsal shape, (d) tongue-tip position. In each figure, three values of the shape factor are shown, -3, 0, and $+3$, and "v.e." indicates the variance explained by the component. Although the remaining three articulators affect the overall shape of the vocal tract, they do not affect tongue shape or location.

The vocal tract shape resulting from a given articulatory configuration is transformed into an area function that is processed to obtain acoustic output and spectral properties of the vocal tract during production of vowels. Software provided by Shinji Maeda, which computes the first three formants corresponding to the vocal tract shape determined by the Maeda articulatory parameters, has been integrated with the DIVA model. These formants are input to Sensyn, a formant synthesizer (Klatt, 1980) provided by Sensimetrics, Inc., which produces speech samples in mu-law format at 8000 samples per second. These samples are written to the audio device (/dev/audio in UNIX) on the Sun workstation.

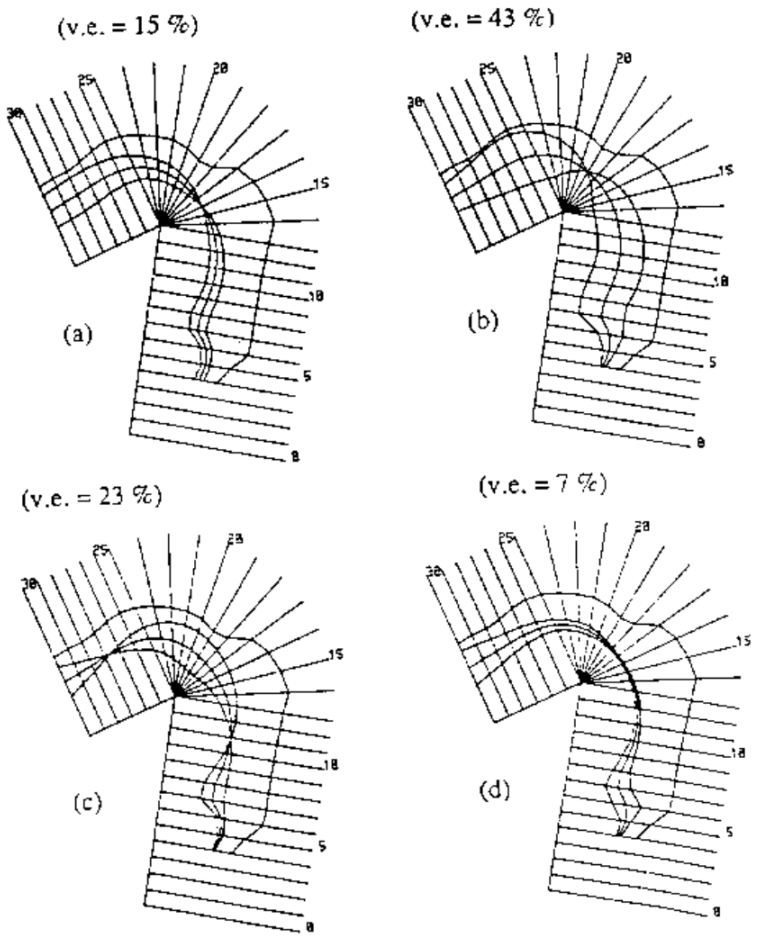


Figure 2·4: Ranges of 4 Maeda articulators and their effect on tongue shape and position. (a) Jaw position, (b) tongue-dorsal position, (c) tongue-dorsal shape, (d) tongue-tip position. In each figure, three values of the shape factor are shown, −3, 0, and +3, and "v.e." indicates the variance explained by the component. (Reprinted from Maeda, 1990, Figure 3.) The remaining 3 Maeda articulators (lip aperture, lip protrusion, and larynx height) do not affect tongue shape or position.

2.2.2 Planning Position Vector

In the present model, the planning space is an acoustic-like space whose dimensions are the formant frequencies F1 and F2 of the acoustic speech signal. In general, the planning space may include many different types of acoustic state information such as F3, amplitudes of F1, F2, and F3, and fundamental frequency. Subsequent chapters explore the hypothesis that the acoustic planning space consists of a wavelet representation of the vowel spectrum.

The Planning Position Vector encodes a scaled representation of the position in the planning space for use in computing the desired direction of movement. In the simulations reported in this chapter, the Planning Position Vector consists of the frequencies of the first two formants, F1 and F2, normalized so that the logarithm of the frequency is scaled to fall within the interval $[0, 1]$, as required by the DIVA model. The Planning Position Vector is defined to be the vector (f_{1+}, f_{1-}, f_{2+}, f_{2-}), where:

$$f_{1+} = \frac{\log(F1) - \log(F1_{\min})}{\log(F1_{\max}) - \log(F1_{\min})} \tag{2.1}$$

$$f_{1-} = 1.0 - f_{1+} \tag{2.2}$$

$$f_{2+} = \frac{\log(F2) - \log(F2_{\min})}{\log(F2_{\max}) - \log(F2_{\min})} \tag{2.3}$$

$$f_{2-} = 1.0 - f_{2+} \tag{2.4}$$

where $F1_{\min}$, $F1_{\max}$, $F2_{\min}$, and $F2_{\max}$, are the minimum and maximum values of $F1$ and $F2$ that can be encountered during babbling. By virtue of this definition of the acoustic state, the sum of the agonist and antagonistic pairs of state variables is 1 for both F1 and F2 and their values are limited to the range [0, 1]. Logarithmic scaling of the frequencies was used to reduce the magnitude of the variation of high frequency

formants relative to low frequency formants. Without this type of normalization, the weights that code the inverse kinematic map are too large, even though the percent change in formant frequency is the same.

2.2.3 Orosensory Feedback and the "Forward Model"

The Planning Position Vector stage represents the current state of the vocal tract within the planning reference frame. This is used to calculate the desired movement direction (in the planning space), which is formed by subtracting the Planning Position Vector from the current sound's target at the Planning Direction Vector stage.

How is the current acoustic state derived by the system? It is hypothesized that humans have access to at least three types of information that convey the state of the vocal tract within the planning reference frame. One source of acoustic information is the auditory system, which can provide formant values from a self-generated acoustic signal. Input from the auditory system is crucial for the acquisition of speech skill. However, an absence of auditory input, resulting, for example, from profound deafness, does not prevent the production of fluent speech if the onset of deafness occurs later in life. (See Section 2.2.4 for more details.) This suggests that sources of information other than the auditory system are used for the computation of the current acoustic state. These sources include the motor command efference copy (a copy of the command sent to the muscles) and *orosensory feedback*, consisting of tactile (i.e., somatosensory) and proprioceptive (muscle spindle) feedback information.

An orosensory dimension (Perkell, 1980; Guenther, 1993, 1994b) is an acoustically important measurement of the vocal tract state or configuration and is derived from tactile or proprioceptive feedback. Perkell (1980) suggested that speech production planning is carried out directly in an orosensory reference frame. However, this thesis assumes that vowel production planning is performed in an acoustic reference frame,

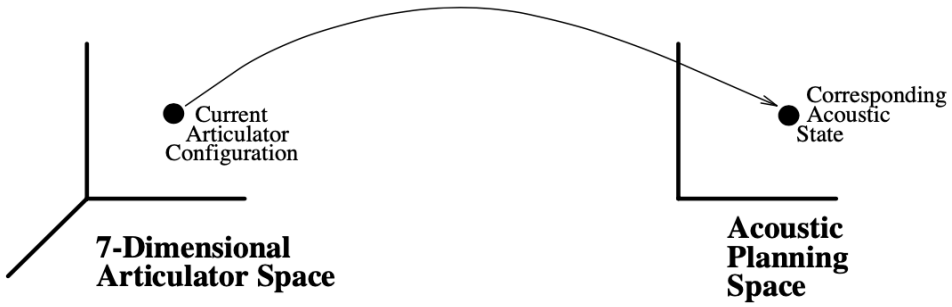


Figure 2·5: Forward Map: Schematic representation of functional mapping from the 7-dimensional articulatory space to the planning (spectral) space.

and that orosensory information is transformed by a neural mapping, or "Forward Model" (e.g., Jordan & Rumelhart, 1992), into the necessary acoustic information. While orosensory information (in the absence of acoustic state information) may play a role in the planning of consonant sounds and may contribute to the planning of vowels, it is assumed that acoustic planning alone is *sufficient* for vowel planning. Furthermore, it is assumed that this efference and orosensory feedback is sufficient to compute the current acoustic state in the absence of auditory input.

In order to determine the relative importance of tactile and proprioceptive feedback for speech production, Borden, Harris, and Oliver (1973) conducted a study in which adults received a bilateral mandibular nerve block. They found that prominent distortions of speech quality occurred only in /s/ clusters. Borden (1979) claims that this interference is with taction, not proprioception. It would seem, then, that taction is not generally necessary for production of speech. Attempts to block proprioceptive feedback showed that, for large jaw changes at least, spindle feedback (proprioception) was necessary. See also Borden (1980).

The tactile and proprioceptive information can be used to form a forward model (Jordan & Rumelhart, 1992) that maps articulator and vocal tract information into the formant values that result from the current shape of the vocal tract. Figure 2·5

illustrates a forward map as an approximation to a vector function from an articulatory space to the corresponding spectral space. The forward model (or map) is schematized by the filled semicircles at the Planning Position Vector block in Figure 2·2.

In subsequent chapters, simulation results in which a radial basis function approximation of the forward model self-organizes during the babbling process are described. In this chapter, however, the forward model is computed by interpolating formant values corresponding to nearby articulatory positions, which have been stored in a lookup table. A generalization of bilinear interpolation (Press, Teukolsky, Vetterling, & Flannery, 1992, page 123) is used. This approach is very efficient and sufficiently accurate for vowel production.

2.2.4 Auditory Processing

The Auditory Processing component is meant to represent human auditory processing in which an acoustic signal is transduced and preprocessed for use by the speech recognition system. The speech recognition system is required only during the babbling phase and is examined in Section 2.2.5. In the present section, the role of auditory input and feedback is discussed in the context of acquisition of speech in human infants.

In order for speech and language to develop, several types of feedback must be provided to the infant. The onset of canonical babbling (i.e., production of repeated syllables such as "baba" and "dada") is a landmark event in the acquisition of speech, occurring in normal infants between $7-10$ months of age (Oller & Eilers, 1988; Smith, Brown-Sweeney, & Stoel-Gammon, 1989; Davis & MacNeilage, 1990), and many studies have shown that auditory input is required for the onset of canonical babbling and the subsequent development of speech. For example, it has been shown that the

onset of babbling is delayed in hearing-impaired infants (Stoel-Gammon & Otomo, 1986; Stoel-Gammon, 1988; Kent, Osberger, Netsell, & Goldschmidt-Hustedde, 1987) and that the inventory of babbled speech sounds is significantly smaller in hearing-impaired infants (Kent et al., 1987; Oller, Eilers, Bull, & Carney, 1985).

Profoundly deaf infants (for whom amplification may provide assistance) and acochlear infants also provide information about the role of auditory feedback in speech development. Lynch, Oller, and Steffens (1989) studied an acochlear male child who had not begun canonical babbling by 27 months of age. The child was presented with about 208 hours of exposure to the Tactaid II (tactile speech device), during which canonical babbling began. This study showed that auditory input is not necessary for the onset of babbling, but that some form of feedback correlated with the speech signal is required.

In addition to auditory input, the ability to produce (self-generated) speech is also crucial to the occurrence of babbling onset and the subsequent acquisition of speech. Locke and Pearson (1990) studied a tracheostomized infant female, who was unable to speak but had normal hearing, and found that after decannulation (removal of the tube) she had "a tenth of the canonical syllables which might be expected in normally developing infants" (page 1).

These data demonstrate a crucial role for auditory input and feedback, and self-generated speech, features which are central to the present model. In these simulations of DIVA, no attempt is made to transduce actual acoustic input. Instead, all input for the speech recognition system is derived directly from the vocal tract shape. The role of the auditory system for speech production is examined in more detail in the following chapters.

Table 2.1: Vowels produced by DIVA with ranges of formant frequencies (in Hz).

Vowel	F1 Range	F2 Range
/IY/ as in "eve"	220 ¡ F1 ¡ 320	2140 ¡ F2 ¡ 2440
/I/ as in "it"	340 ¡ F1 ¡ 440	1890 ¡ F2 ¡ 2090
/E/ as in "met"	480 ¡ F1 ¡ 580	1740 ¡ F2 ¡ 1940
/AE/ as in "at"	610 ¡ F1 ¡ 710	1620 ¡ F2 ¡ 1820
/A/ as in "father"	630 ¡ F1 ¡ 810	990 ¡ F2 ¡ 1190
/OH/ as in "all"	520 ¡ F1 ¡ 620	740 ¡ F2 ¡ 940
/U/ as in "foot"	390 ¡ F1 ¡ 470	920 ¡ F2 ¡ 1120
/OO/ as in "boot"	250 ¡ F1 ¡ 350	770 ¡ F2 ¡ 970
/UH/ as in "up"	490 ¡ F1 ¡ 570	1090 ¡ F2 ¡ 1260
/ER/ as in "bird"	490 ¡ F1 ¡ 540	1280 ¡ F2 ¡ 1450

2.2.5 Speech Recognition System

The Speech Recognition System is a simple expert system whose input is the Planning Position Vector and whose output is the identity of the current phoneme. Table 2.1 shows the allowed formant ranges for the vowels produced by DIVA. Ranges for the formants are implemented in the form of "if-then" rules for each vowel in the recognition system. Referring to Table 2.1, the rule for recognizing the vowel /IY/, for example, is simply:

$$\text{If } (220 < F1 < 320) \text{ and } (2140 < F2 < 2440), \text{ then /IY/ is being pronounced.}$$

If these conditions are met during babbling, the model activates the cell corresponding to /IY/ in the speech sound map, and after many productions this cell effectively learns a rectangular region target encompassing these ranges. After babbling, activation of this cell causes the vocal tract to move so as to achieve formant values within these ranges. While the model computes the third formant, F3, and uses it to drive the speech synthesizer, vowel identity does not depend on F3, and F3 is not used during learning.

2.2.6 Speech Sound Map

The Speech Sound Map is a layer of cells whose activity determines either the phoneme to be produced (during performance) or the most recently recognized phoneme (during babbling). Only one phoneme may be active at a time. 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 the value 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:

$$s_i = \begin{cases} 1 & : \text{ if recognition system hears } i\text{th sound} \\ 0 & : \text{ otherwise} \end{cases} \tag{2.5}$$

SSM Activities, Performance Phase:

$$s_i = \begin{cases} 1 & : \text{ if production of } i\text{th sound is desired} \\ 0 & : \text{ otherwise} \end{cases} \tag{2.6}$$

where s_i is the activity of the cell corresponding to the ith sound, and the index i takes on a value between 1 and 10, corresponding to the vowels learned by the model.

2.2.7 Planning Direction Vector

This vector consists of a layer of cells which code the direction in which the system moves in order to reach the phonemic target in acoustic space. The value of this vector is the difference between the nominal acoustic target corresponding to the desired phoneme and the current Planning Position Vector. In the simulations reported here,

four cells were used in the Planning Direction Vector. The activities of the Planning Direction Vector cells are governed by the following equations:

Planning Direction Vector cell Activities, Babbling & Performance Phases:

$$d_{j+} = \sum_i s_i z_{ij+} - f_{j+} \tag{2.7}$$

$$d_{j-} = \sum_i s_i z_{ij-} - f_{j-} \tag{2.8}$$

where d_{j+} and d_{j-} are the antagonistically paired Planning Direction Vector cell activities corresponding to the jth acoustic dimension, f_{j+} and f_{j-}, defined in Equations 2.1 - 2.4, are antagonistically paired acoustic feedback signals coding position along the jth dimension of acoustic space, s_i is the activity of the ith Speech Sound Map cell, z_{ij+} is the synaptic weight of the pathway from the ith Speech Sound Map cell to the $j+$th Planning Direction Vector cell, and z_{ij-} is the synaptic weight of the pathway from the ith Speech Sound Map cell to the $j-$th Planning Direction Vector cell. The weights z_{ij+} and z_{ij-} constitute the phonetic-to-acoustic mapping, and depend on the current position in articulatory space.

These equations show that Planning Direction Vector cells receive inhibitory tactile and proprioceptive feedback about the state of the vocal tract, represented by the values f_{j+} and f_{j-}. Planning Direction Vector cells also receive excitatory input via the learned phonetic-to-acoustic mapping; this can be seen as the $\sum s_i z_{ij+}$ and $\sum s_i z_{ij-}$ terms in Equations 2.7 and 2.8. When a cell in the Speech Sound Map is activated for performance of the corresponding sound, this input to the Planning Direction Vector acts as a target in acoustic space for producing that sound. The Planning Direction Vector then represents the difference between the learned acoustic target for the desired sound and the current configuration; this value specifies a desired movement direction in acoustic space that is then mapped into a set of

articulator directions to move the vocal tract in this direction.

The acoustic-to-articulatory mapping is learned during the first stage of babbling (see Section 2.3.1 for details). During this stage, phoneme recognition is not performed. Therefore, all s_i are set to zero and no excitatory input propagates to the Planning Direction Vector cells. But changes in the configuration of the vocal tract cause changes in the Planning Direction Vector activities, by virtue of the feedback signals f_{j+} and f_{j-}. These changes drive learning in the acoustic-to-articulatory mapping. During the second babbling stage, random production of a speech sound will result in activation of the corresponding s_i. Planning Direction Vector cell activity now reflects the difference between the current vocal tract configuration (from the f_{j+} and f_{j-}) and the acoustic target for that speech sound (encoded by the weights z_{ij+} and z_{ij-}). This difference drives learning in the phonetic-to-acoustic mapping (see Section 2.3.2 for details).

2.2.8 Articulator Direction Vector

The Articulator Direction Vector consists of a set of cells that command movements of the articulators. These cells code the movement direction in articulatory space which is predicted to reduce the acoustic distance to the phoneme target. 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. During babbling, Articulator Direction Vector cells are activated to produce movements of the articulators. It is assumed that this occurs via an endogenous random generator that overrides other Articulator Direction Vector inputs during babbling (see Bullock et al., 1993; Gaudiano & Grossberg, 1991). During performance, the endogenous random generator is disabled, and Articulator Direc-

tion Vector cells are activated by excitatory input from the acoustic-to-articulatory mapping. Specifically, Planning Direction Vector cell activities are governed by the following equations:

Articulator Direction Vector Activities, Babbling Phase:

$$a_{k+} = \begin{cases} 1 & : \text{ when endogenously activated} \\ 0 & : \text{ otherwise} \end{cases} \tag{2.9}$$

$$a_{k-} = \begin{cases} 1 & : \text{ when endogenously activated} \\ 0 & : \text{ otherwise} \end{cases} \tag{2.10}$$

Articulator Direction Vector Activities, Performance Phase:

$$a_{k+} = \sum_j [d_{j+}]^+ w_{j+k+} + \sum_j [d_{j-}]^+ w_{j-k+} \tag{2.11}$$

$$a_{k-} = \sum_j [d_{j+}]^+ w_{j+k-} + \sum_j [d_{j-}]^+ w_{j-k-} \tag{2.12}$$

where a_{k+} and a_{k-} are the antagonistic pair of activities corresponding to the kth articulatory degree-of-freedom, w_{j+k+} is the synaptic weight projecting from the j+th Planning Direction Vector cell to the k+th Articulator Direction Vector 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 acoustic-to-articulatory mapping. The acoustic-to-articulatory mapping is discussed in detail in Section 2.3.1.

2.2.9 Phoneme String

This component is a string of one or more phoneme identifiers to be produced by the model. It is hypothesized to originate in "higher cognitive centers". These strings are input to the simulation by the user.

2.2.10 GO Signal

The GO signal (Bullock & Grossberg, 1988) is used to multiplicatively gate the movement commands at the Articulator Direction 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 humans. The equation governing articulator directions is as follows:

Articulator Velocities:

$$v_k = G \times [a_{k+} - a_{k-}] \tag{2.13}$$

where v_k is the velocity along the kth 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, and is fixed at the value of 1.0 during the performance phase. The issue of velocity control is not otherwise treated explicitly in this dissertation.

2.3 Learned Mappings

This section discusses the learning of the Acoustic-to-Articulator mapping, represented by filled semicircles in Figure 2·2 between the Planning Direction Vector and the Articulator Direction Vector, and the Phoneme-to-Acoustic mapping, represented by filled semicircles in Figure 2·2 between the Speech Sound Map and the Planning Direction Vector. 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. Acquisition of speaking skills in DIVA consists of finding appropriate parameters, or synaptic weights, for the acoustic-to-articulatory and phonetic-to-acoustic mappings during the two stages of the babbling phase. Details of these mappings and the learning processes involved

during babbling are described in the following paragraphs.

2.3.1 The Acoustic-to-Articulatory Mapping

In order to compute the Acoustic-to-Articulatory Mapping, first consider the forward model (Section 2.2.3) which maps articulatory configurations to their corresponding points in acoustic space. (A schematization of the forward map is shown in Figure 2·5.) Assume, in general, M articulators and N planning space dimensions, and let θ be the vector of articulator values, i.e., $\theta = [\theta_1, \cdots, \theta_M]^T$, where θ_i is the ith Maeda articulator value, and let x be the vector of planning variables, i.e., $x = [x_1, \cdots, x_N]^T$. Then the forward model is defined to be the set of functions $(f_1, \cdots, f_N)$ such that

$$\begin{cases} x_1 &= f_1(\theta_1, \cdots, \theta_M) \\ \vdots & \\ x_N &= f_N(\theta_1, \cdots, \theta_M) \end{cases} \tag{2.14}$$

The set of equations defined by 2.14 can be regarded as a vector equation

$$x = f(\theta) \tag{2.15}$$

where f is the vector function $f = (f_1, \cdots, f_N)$.

In general, $J(\theta)$ is not constant. Assume that the current articulatory configuration is given by the vector θ_0. The acoustic-to-articulatory mapping at this articulatory configuration is found by differentiating Equation 2.15, i.e.,

$$dx = f'(\theta)d\theta = J(\theta)d\theta \tag{2.16}$$

where $J(\theta)$ is the Jacobian matrix associated with the function f, whose elements J_{ij} are the partial derivatives $\frac{\partial f_i}{\partial \theta_j}\Big|_{\theta_0}$. Given a desired movement dx in the acoustic space,

it is desired to compute the corresponding movement $d\theta$ in articulatory space, which can be obtained by solving Equation 2.16 for $d\theta$. When J is not a square matrix, the inverse of J does not exist, and it is necessary to compute the pseudoinverse of J. Let this pseudoinverse be denoted by $J^{-1}(\theta)$. Then

$$d\theta = J^{-1}(\theta)dx \tag{2.17}$$

The vectors dx and $d\theta$ are related to, but not equal to, the Planning Direction Vector and the Articulator Direction Vector, respectively. The latter are neural network constructs and employ an agonist and antagonist pair of components for each component of the former vector. The matrix form of Equation 2.17 is given by

$$\begin{bmatrix} d\theta_1 \\ \vdots \\ d\theta_m \end{bmatrix} = \begin{bmatrix} J_{11}^{-1} & \cdots & J_{1n}^{-1} \\ & \ddots & \\ J_{m1}^{-1} & \cdots & J_{mn}^{-1} \end{bmatrix}_{m \times n} \begin{bmatrix} dx_1 \\ \vdots \\ dx_n \end{bmatrix} \tag{2.18}$$

For a seven articulator system and an acoustic space consisting of two formants, Equation 2.18 reduces to

$$\begin{bmatrix} d\theta_1 \\ \vdots \\ d\theta_7 \end{bmatrix} = \begin{bmatrix} J_{11}^{-1} & J_{12}^{-1} \\ \vdots & \vdots \\ J_{71}^{-1} & J_{72}^{-1} \end{bmatrix}_{7 \times 2} \begin{bmatrix} dx_1 \\ dx_2 \end{bmatrix} \tag{2.19}$$

The desired value of $d\theta_i$ is then given by

$$d\theta_i = \sum_j J_{ij}^{-1}(\theta)dx_j \tag{2.20}$$

The transformation performed by the acoustic-to-articulatory mapping defined in Equation 2.18 is an inverse kinematic mapping and can be envisioned as a transfor-

mation of the movement specification from a sensory coordinate frame to a motor coordinate frame. As described in Section 2.2.7, the Planning Direction Vector cells code the direction from the current vocal tract configuration to the target region in acoustic space. Multiplying this vector by the matrix of weights in the acoustic-to-articulatory pathways (Equations 2.11 and 2.12) effectively produces a vector describing the direction of desired movement in the motor coordinate frame. This vector serves as the basis for commanded directions of the articulators.

In the first stage of babbling, the DIVA model learns a mapping from directions in acoustic space (coded by the Planning Direction Vector) to movement directions in articulator space (coded by the Articulator Direction Vector). A portion of this mapping is shown in Figure 2·6. Learning of the acoustic-to-articulatory mapping occurs as follows. Randomly activated Articulator Velocity Vector cells cause movements of the speech articulators which are reflected through acoustic feedback as changes in activity of the Planning Direction Vector cells. It is these *changes* in Planning Direction Vector activities, rather than the magnitude of activities, that drive learning in the acoustic-to-articulatory pathways according to the following equations:

Acoustic-to-Articulatory Mapping Learning Equations:

$$\frac{d}{dt}w_{j+k+} = -\varepsilon_1(\frac{d}{dt}d_{j+})(a_{k+} + \alpha\sum_i w_{ik+}(\frac{d}{dt}d_i)) \tag{2.21}$$

$$\frac{d}{dt}w_{j+k-} = -\varepsilon_1(\frac{d}{dt}d_{j+})(a_{k-} + \alpha\sum_i w_{ik-}(\frac{d}{dt}d_i)) \tag{2.22}$$

$$\frac{d}{dt}w_{j-k+} = -\varepsilon_1(\frac{d}{dt}d_{j-})(a_{k+} + \alpha\sum_i w_{ik+}(\frac{d}{dt}d_i)) \tag{2.23}$$

$$\frac{d}{dt}w_{j-k-} = -\varepsilon_1(\frac{d}{dt}d_{j-})(a_{k-} + \alpha\sum_i w_{ik-}(\frac{d}{dt}d_i)) \tag{2.24}$$

where ε_1 is a velocity learning rate parameter and α is a velocity training ratio, and

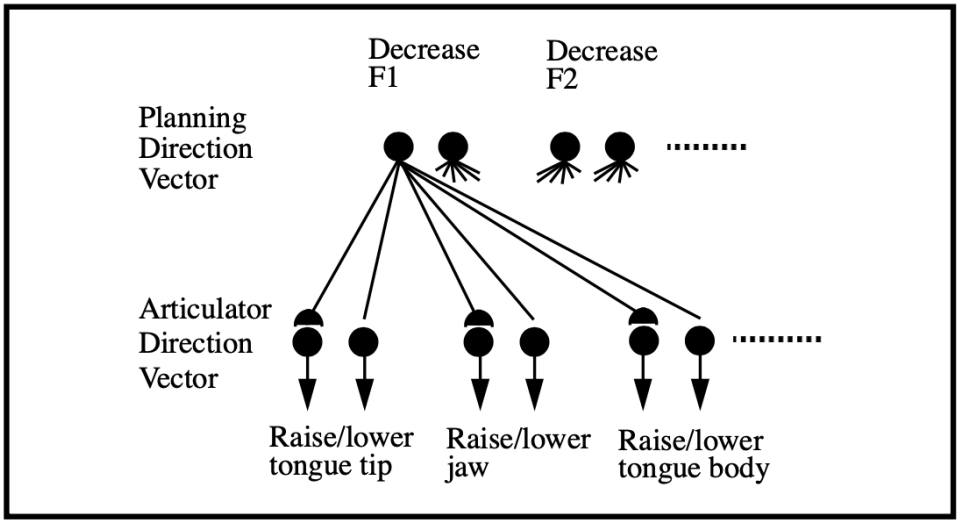


Figure 2·6: Inverse map: Portion of the acoustic-to-articulatory mapping after babbling. Planning Direction Vector cells, each coding a desired movement direction in Acoustic space, project with large weights to Articulator Direction Vector cells that move the vocal tract in the appropriate direction. Projections to other Articulator Direction Vector cells have withered away to zero during learning. Activity at a Planning Direction Vector 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.

where i can take on the values of $k+$ and $k-$ for all possible values of k. In the simulations, $\alpha = 10.0$ and $\varepsilon_1 = 500.0$. Thus, a decrease in a Planning Direction Vector cell's activity results in an increase in the weight projecting from the Planning Direction Vector cell to active Articulator Direction Vector cells; these Articulator Direction Vector cells are responsible for the movements that resulted in the initial decrease of Planning Direction Vector cell activity. In this way, each Planning Direction Vector cell learns a set of articulator directions that cause movements to decrease the Planning Direction Vector cell's activity, i.e., articulator movements that move the vocal tract in the desired direction.

Guenther (1992) and Bullock, Grossberg, and Guenther (1993) discuss how the inverse kinematic mapping described by Equations 2.21 - 2.24 can lead to desir-

able motor-equivalent behavior, and they claim that the inverse kinematic mapping roughly approximates the Moore-Penrose pseudo-inverse of the nonlinear forward map.

The number of weights required to encode the acoustic-to-articulatory mapping is 3024, or 4 (F1 and F2, increase and decrease) times 14 (7 Maeda articulatory directions, increase and decrease) times 54 tessellated regions.

2.3.2 The Phonetic-to-Acoustic Mapping

The synaptic weights in the pathways projecting from a Speech Sound Map cell to the Planning Direction Vector cells represent an acoustic target for the corresponding speech sound in acoustic space. When the changing vocal tract configuration gives rise to an acoustic state that is identified by the Speech Recognition system as corresponding to 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-acoustic pathways projecting from that cell, and, as described in the following paragraphs, this allows the model to modify the acoustic target for the speech sound based on the current configuration of the vocal tract as seen through orosensory feedback at the Planning Direction Vector stage.

The neural mechanism used to learn the rectangular 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:

Phonetic-to-Acoustic Mapping Learning Equations:

$$\frac{d}{dt}z_{ij+} = \varepsilon_2 s_i(\alpha_2 z_{ij+} - [d_{j+}]^+) \tag{2.25}$$

$$\frac{d}{dt}z_{ij-} = \varepsilon_2 s_i(\alpha_2 z_{ij-} - [d_{j-}]^+) \tag{2.26}$$

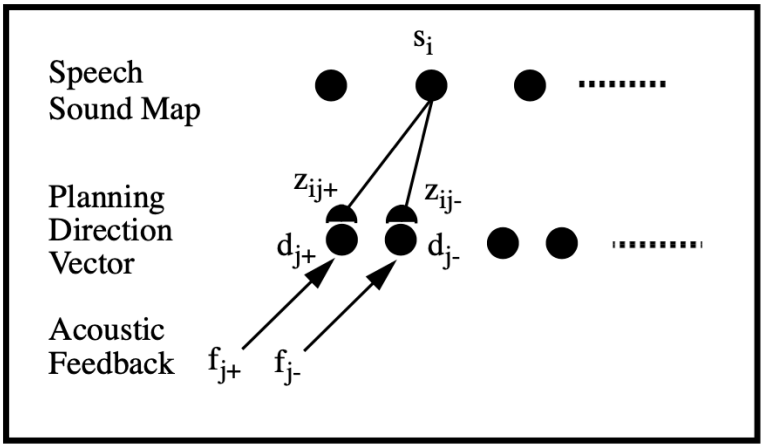


Figure 2·7: Target map: Portion of the phonetic-to-acoustic mapping from a Speech Sound Map cell to the antagonistic pair coding one dimension of acoustic space.

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 2.25 and 2.26 ensure that modification of a given phoneme's acoustic 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 acoustic rectangular region target. Target regions in DIVA are learned in a manner which is similar to the learning of rectangular categories in Fuzzy ART (Carpenter, Grossberg, & Rosen, 1991a).

Figure 2·7 schematizes the mapping from a Speech Sound Map cell to the antagonistic pair coding one dimension of the Planning Direction Vector. The acoustic feedback signal antagonistic pairs (f_{j+},f_{j-}) each sum to a constant value of 1. Assume a large value of ε_2 and a very small value of α_2 in Equations 2.25 and 2.26. In the simulations, $\varepsilon_2 = 4.0$ and $\alpha_2 = 0.0$. 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 2.11 and 2.12 it is clear that

during performance only positive d_{j+} and d_{j-} will activate articulator movements. With $(z_{ij+}, z_{ij-}) = (0.4, 0.6)$, from Equations 2.7 and 2.8 it is seen 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-}).

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, it is seen from Equations 2.7 and 2.8 that a positive d_{j+} or d_{j-} will only result if (f_{j+}, f_{j-}) is outside the range $(0.4 \leq f_{j+} \leq 0.5, 0.5 \leq f_{j-} \leq 0.6)$. This range thus defines a rectangular region attractor. Further decreases in the weight values will result in further increases in the size of the rectangular region attractor. The target region converges to an approximation of the vowel category region encoded by the speech recognition system, and learning is stable as long as there is no overlap of the vowel categories in formant space.

The number of weights required to encode the phonetic-to-acoustic mapping is 40, or 4 (F1 and F2, increase and decrease) times 10 vowels. The number of weights for this mapping is significantly smaller than the number of weights required for the acoustic-to-articulatory mapping.

2.3.3 Tessellation of Articulator Space

The acoustic-to-articulatory mapping varies with articulatory configuration. The physics of the vocal tract determines the effect of each articulator on the resulting formant values in different vocal tract configurations. This property of the vocal tract requires that the model learn a different inverse kinematic mapping at each of many vocal tract configurations. For example, when the tongue is in a front position, a movement in larynx height has little or no effect on the formant values. Therefore, the corresponding elements of the inverse Jacobian matrix are zero. However, the

same larynx height change when the tongue is far back has a significant effect on F1, and the corresponding element of the inverse Jacobian matrix is relatively larger. It is suggested that humans learn these differences and use them during the production of vowels.

In order to encode these differences in the approximate inverse Jacobian learned by the model, the articulator space is partitioned or tessellated in a simple manner. Each articulatory dimension is divided into a small number of equal-length intervals. Taken together, these articulator regions define hyper-rectangles in articulatory space. The acoustic-to-articulatory mapping for each tessellated region is learned by the model during the babbling phase. During learning, and later during performance, the Articulator Position Vector determines the tessellation region and the acoustic-to-articulatory mapping corresponding to that region is loaded and used.

The choice of the number and size of tessellated regions used in simulations was hand-crafted to produce smooth movement between all vowel targets. Acceptable vowel production was obtained with as few as 54 tessellation regions of equal size. Using a larger number of regions makes the computation of the acoustic-to-articulatory mapping more accurate, but it also increases the time required for babbling and the memory used to store the mappings. Although the boundaries of the tessellated regions are not learned in this model, the weights which code the acoustic-to-articulatory mapping *are* learned. These weights enable the model to adapt to varying articulatory systems. Tessellated regions were used for simplicity in the model, but resulted in numerous simulations designed only to choose the best tessellation. To avoid this inconvenience, Guenther et al. (1998) employ hyperplane radial basis functions (HRBFs) to approximate the acoustic-to-articulatory mapping. HRBFs were also used in subsequent chapters of this dissertation in the wavelet auditory version

of DIVA. The role of a single HRBF is analogous to a tessellated region, but provides better smoothing between the HRBFs, and can adaptively self-organize during babbling.

The computational details of the babbling phase have been described above. However, a few words need to be said about the role of random babbling and the existence of tessellation regions in the acoustic-to-articulatory mapping. While it is expected that fully random babbling can be used to learn the acoustic-to-articulatory mapping, a more structured approach to setting the articulatory direction vector activities was used during this babbling subphase. This was necessary to ensure that each tessellation region learned its inverse mapping in a reasonable time, and also because the emphasis of this research was not on babbling *per se*.

To ensure that each tessellation region learned its inverse mapping, the following algorithm was used to set the Articulator Direction Vector activities. For each tessellation region, the center of the region was located. If this position in articulatory space corresponded to a place where the vocal tract was closed, then a place within the region closer to the neutral vocal tract configuration was chosen. Then each articulator was activated in sequence and its effect was learned. Because of the form of the learning law, for slow learning, the values of the learned weights are independent of one another, thereby enabling this approach. This process was repeated for each articulator until the weight values equilibrated, and the next tessellation region was then chosen. To ensure correct learning of these weights, it was necessary to make the movements of the articulators very small during this babbling stage (to ensure that the vocal tract configuration remained within the current tessellated region).

Since it is not necessary to tessellate the acoustic space to learn an accurate phoneme-to-acoustic mapping, random babbling was used during the second subphase

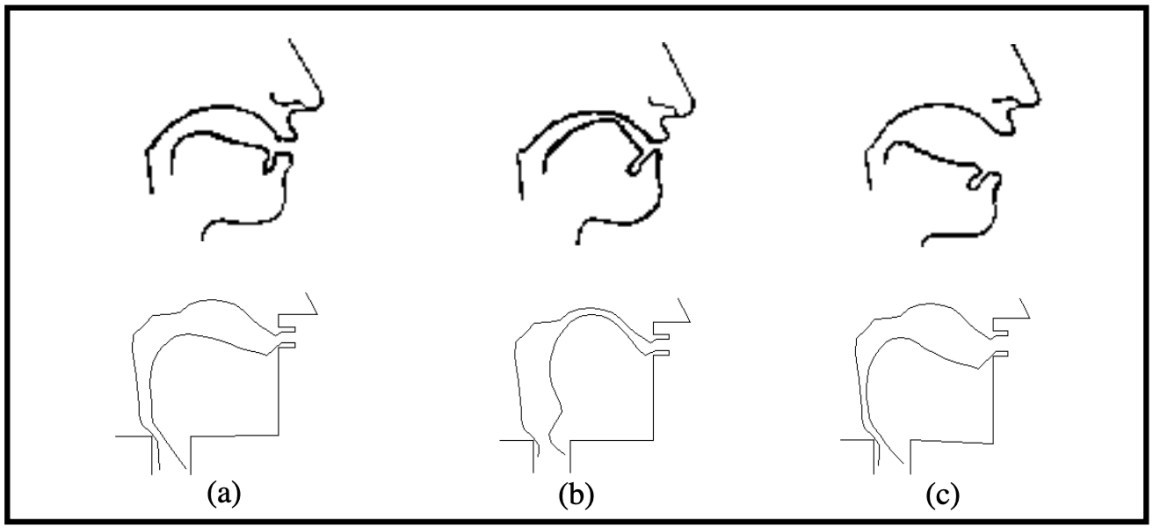


Figure 2·8: Vocal tract configurations corresponding to different vowels. The top row shows schematics of the profiles used by humans (top row; after Flanagan, 1972) and the bottom row shows the configurations produced by the model. These configurations occur despite the fact that no explicit vocal tract shape targets are used in the model. (a) The central vowel /UH/ as in "up". (b) The high front vowel /IY/ as in "beet". (c) The low back vowel /A/ as in "father".

of the babbling.

2.4 Computer Simulation Results

Simulations of the model reported in this chapter were run on a Sun Microsystems SPARC-10 running SunOS 4.1.3. The X Window System (X11R5) with Motif was used to generate the graphical output. The software was written in the C programming language, compiled and linked with gcc, and optimized and profiled to maximize run-time performance.

2.4.1 Production of Vowels

Ten English vowels were learned during babbling. Synthesis of the model's vocal tract configurations while producing each vowel in isolation resulted in vowel sounds that

were intelligible to the author. Each vowel can be produced by the model from any starting configuration of the vocal tract. As illustrated in Figure 2·8, the resulting vocal tract shapes correspond roughly to shapes seen in humans producing the same vowels, even though no vocal tract shape information is explicitly encoded in the targets learned by the model.

2.4.2 Acoustic Planning Enhances Motor Equivalent Speech Production

Perturbation studies (e.g., Abbs, 1986; Abbs & Gracco, 1984; Lindblom, Lubker, & Gay, 1979; Savariaux et al., 1995) have established that the speech production system exhibits motor equivalence by using new articulator configurations that preserve the perceptual identity of a phoneme when the default articulator configuration for that phoneme is not possible due to an externally imposed constraint such as a bite block or lip tube. The present model, which utilizes acoustic space planning, assumes that speaker compensation is geared toward maintaining the relevant auditory perceptual aspects (in this case, formants) of the sound being produced. Moreover, it is assumed that invariant constriction targets would unnecessarily limit the motor equivalent capabilities of the speech production system. One reason for this is that the number of parameters needed to specify vocal tract shape and glottal source (non-acoustic) characteristics is much larger than the number of parameters needed to specify the acoustic target, thereby placing an unnecessary burden on the vowel production system. In addition, it is possible that the presence of obstructions (food, bite blocks, etc.) may affect the acoustic output in ways other than by merely modifying the vocal tract shape. Without detailed knowledge of these interactions, the vowel production system is unable to accurately produce the required vowel by planning only for vocal tract shape. On the other hand, employing an acoustic target is much simpler and is more likely to result in production of the desired vowel with the relevant audi-

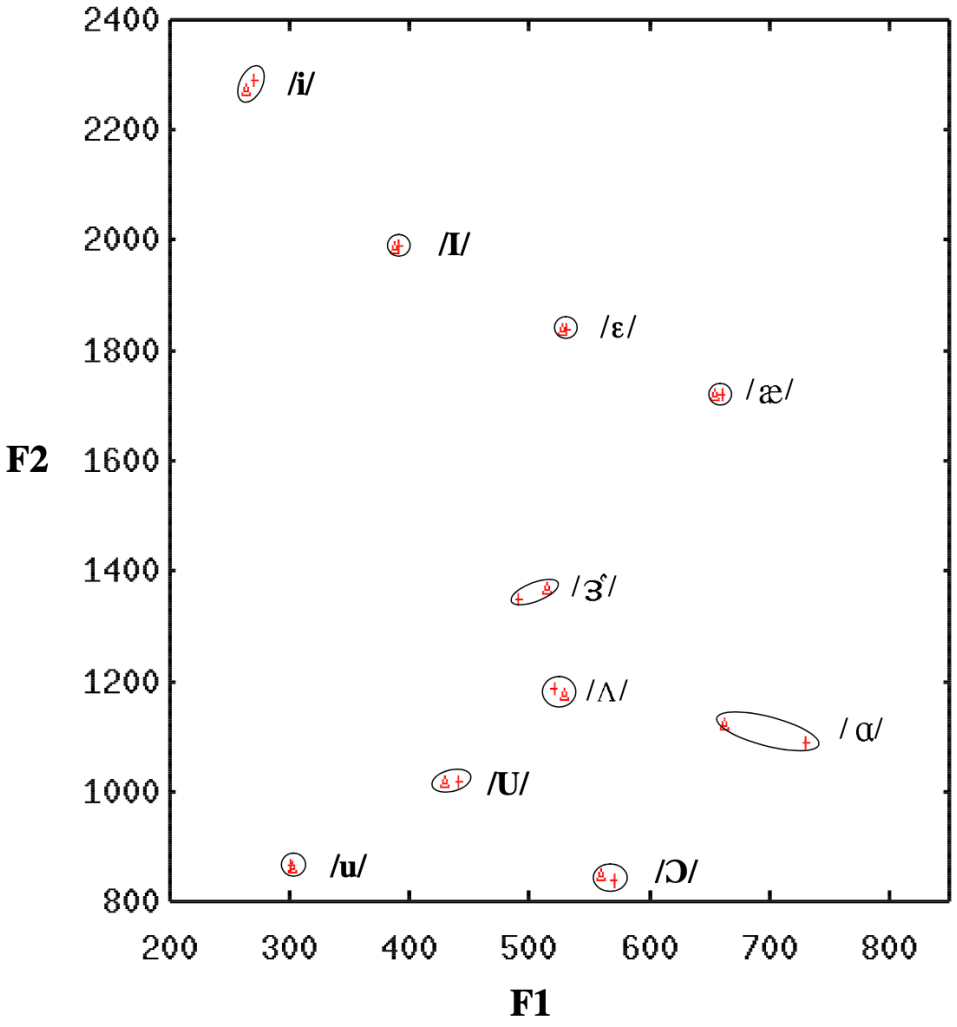


Figure 2·9: Typical values of F1 and F2 for American English vowels (indicated by crosses) and corresponding values produced by the model (indicated by triangles) when starting from a neutral vocal tract configuration. An ellipse is drawn around the typical value and the value produced by the model for each vowel. (Reprinted from Guenther, Hampson, & Johnson, 1998.)

tory perceptual features. This higher level of motor equivalence, made possible by the acoustic target strategy, would clearly be advantageous to the vowel production system since the goal of speech is ultimately to produce recognizable phonemes.

The following simulations illustrate the motor equivalent capabilities of the DIVA model when using acoustic space targets for vowels (Johnson & Guenther, 1995). Figures 2·9 and 2·10 show the results of simulations carried out with the current model. Figure 2·9 illustrates the model's performance in the absence of constraints on the articulators. Typical values of F1 and F2 for the vowels are indicated by crosses, and values produced by the model when starting from a neutral vocal tract configuration are indicated by triangles. An ellipse is drawn around the typical value and the value produced by the model for each vowel. The model gets very close to the target for each vowel, although the /A/ produced by the model is significantly farther away from its target value than the other vowels. (This latter result appears to reflect a difficulty inherent to the Maeda vocal tract in reaching the typical /A/ formants specified by Rabiner and Schafer, 1978.) Figure 2·10 illustrates the model's performance when the jaw is fixed in a position that is unnatural for most of the vowels, as would occur if a bite block were held between the teeth. Rectangles indicate formant values that would arise without compensation for the new jaw position, and lines connect these values to the typical values for the corresponding vowels. Despite the large formant shift induced by the bite block, the formant values produced by the model in the bite block condition are nearly identical to values produced in the unconstrained condition (Figure 2·9), indicating full compensation for the bite block.

This compensation occurs automatically in the model, even though no training is performed with the jaw constraint present. Load compensation in the model can be understood by considering Figure 2·6 in which each planning direction vector cell has

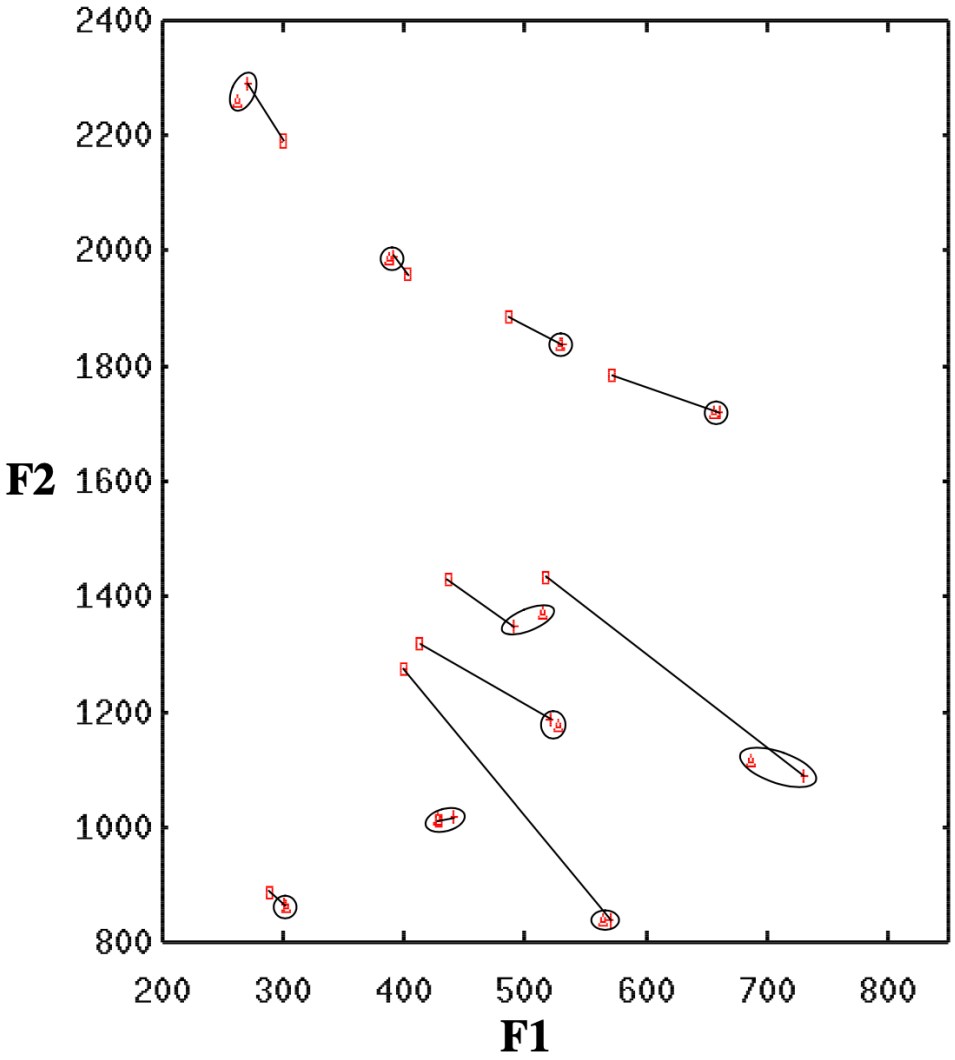


Figure 2·10: "Bite block" simulation in which the jaw parameter is clamped at a value of 0. Crosses indicate typical formant values for male speakers, triangles indicate the formant values produced by the model, and ellipses are drawn around corresponding typical values and model values. Rectangles indicate formant values that would arise without compensation. The formant values produced by the model in this condition are nearly identical to values produced in the unconstrained condition (see previous figure), indicating full compensation for the bite block. This compensation occurs automatically in the model, even though no training was performed with the jaw constraint present. (Reprinted from Guenther, Hampson, & Johnson, 1998.)

a weighted projection to every articulator direction vector cell capable of bringing about the desired movement in the planning direction. The value of the weight codes the degree to which the articulatory movement can bring about the desired change in the planning space. For example, a weight of zero (or a missing connection) indicates that the articulatory movement cannot bring about the desired change in the planning space, a weight of w implies a fixed amount of change in the planning space for a given movement of the corresponding articulator, and a weight of $2w$ means that twice as much change in the planning space occurs for the same articulatory movement. During normal operation, a desired movement in the planning space projects to its articulators, and each of these articulators contributes a portion of the movement needed to reach the acoustic target. A load can be modeled as a blocked articulator, i.e., an articulator that is unable to move. When a load is applied, the other articulators (which are not blocked) continue to move the system closer to the acoustic target. As long as there are unblocked articulators with nonzero weights projecting from the desired planning direction vector cell, the desired movement in the planning space will occur, even in the presence of one or more blocked articulators. Load compensation in the model is simulated by simply setting the relevant articulatory parameter to a fixed value (usually zero) which is then not allowed to change during the production of the phoneme.

2.5 Discussion and Conclusions

The results presented in this chapter demonstrate that acoustic planning using formant frequencies is feasible in DIVA, and that the highly nonlinear inverse kinematic map does not prevent smooth trajectory formation in the planning space. Smooth trajectories in formant space were obtained in spite of the fact that a discontinuous

inverse kinematic mapping was learned (i.e., multiple tessellated regions of equal size in articulatory space with fixed inverse map within each region). If fixed inverse maps in adjacent regions were sufficiently different, then movements from one region to the other would undergo a velocity or direction discontinuity at the boundary. The fact that this was not observed suggests that the inverse maps in adjacent tessellated regions were not significantly different. Although simulations reported in later chapters, and in Guenther et al. (1998), use RBFs which learn smoother approximations to these kinematic maps than is possible using tessellations, and which are able to self-organize during babbling, by careful tessellation of articulatory space, possible singularities in the inverse map (locations at which zero movement vectors result) were either avoided or eliminated entirely. By virtue of the Maeda articulatory system, the present model explains a wider array of data than the earlier model. The ten model vowels were correctly produced from a neutral vocal tract starting position, and acoustic output corresponding to these vowel productions were easily recognized by the author during informal listening tests. In addition, the model uses vocal tract shapes similar to humans, even though vocal tract shape information is not explicitly coded in the vowel targets. Furthermore, the motor-equivalence properties of the original model are preserved and enhanced by using acoustic planning. Each of the ten vowels were successfully produced with the jaw blocked at various positions, demonstrating motor equivalence. With the jaw blocked, other articulators such as the tongue compensated, allowing the vocal tract to assume an overall shape that reached the acoustic target for the vowel. Phonemes produced with the jaw blocked were acoustically indistinguishable from phonemes that were produced with an unconstrained jaw.

Chapter 3

A Wavelet Auditory Representation of Acoustic Spectra for Vowel Perception and Production

The simulation results reported in Chapter 2 showed that a simple acoustic planning space based on formant frequencies is sufficient for production of vowel sounds. In this chapter, evidence from vowel perception experiments is examined which justifies questioning the role of formants and supports a dominant role for gross spectral shape in speech. Then data are presented that suggest that primary auditory cortex codes vowel spectra using a multiscale, wavelet-like, representation. Inspired by these physiological results, the wavelet auditory representation of vowel spectra is then proposed. This representation is utilized to explain the spectral center of gravity effect, in which nearby spectral peaks are averaged into a single peak. The next chapter presents simulation results demonstrating that vowel movement planning based on the proposed wavelet auditory representation successfully produces each of the target vowels using vocal tract configurations similar to humans. Moreover, the wavelet auditory representation is easier to compute than formants.

3.1 Gross Spectral Shape Versus Formants

Are formant frequencies the relevant parameters for the control of speech movements? Traditionally, formants (prominent spectral peaks corresponding to vocal tract resonances) and ratios of formants have been used as a low-dimensional characterization

of the vowel spectrum, and they have played a prominent role in theories of vowel perception and production. But a number of problems exist with theories based on formant representations of vowel spectra.

One problem with formant-based theories of vowel perception and production is that there is no direct evidence that the human brain carries out the necessary computations to extract formant parameters from the speech spectrum. On the other hand, the work of Shamma and his colleagues (reviewed in Section 3.3) have demonstrated the existence of cells in primary auditory cortex that have wavelet-like receptive fields and may serve as a basis for the representation of the log magnitude spectrum of sounds (Yang, Wang, & Shamma, 1992; Wang & Shamma, 1994a, 1995). Their research provides direct evidence that the spectrum of a vowel is represented in primary auditory cortex in a manner that utilizes multiple spectral scales. In particular, Shamma has found that the brain uses a multiscale representation of the acoustic spectrum at every frequency in auditory cortex. By definition, a *multiscale representation* of a function is one in which the largest scales of the representation form a rough approximation of the function, and in which the approximation may be refined by adding progressively smaller scales. Multiscale (and related multiresolution) approximations are discussed by Mallat (1989).

The rough approximation of the vowel spectrum, using only the largest spectral scales, constitutes the *gross spectral shape* of the vowel spectrum. A number of researchers have suggested that gross spectral shape may be better correlated with vowel perception than formants. For example, Zahorian and Jagharghi (1993) write: "...it is not clear that the formants play a fundamental role in speech perception" (page 1966). They go on to show that a representation of vowel spectra based on gross spectral shape predicts vowel identity better than formants. Miller (1989), who

proposed a formant-ratio representation of vowel spectra, writes: "Clearly, spectral shapes are highly correlated with the location of formant frequencies and, therefore, such locations offer an attractive approach to the specification of spectral shape. However, it would not be surprising if related metrics based on the entire spectral shape will be required in the future as it is well known that 'formant tracking' in continuous speech is quite difficult" (page 2132). Because spectral peaks are correlated with vowel identity, it is certainly true that a representation of gross spectral shape (that preserves the spectral peaks) must also be correlated with vowel identity. However, not all representations of gross spectral shape are equally good for predicting vowel identity, and the next section examines several representations of gross spectral shape.

Bladon (1982) presented arguments for gross spectral shape in vowel perception. The first of Bladon's reasons is that changes in formants also change spectral shape. Thus, manipulations of formants in psychophysical experiments are also affecting the gross spectral shapes of the stimuli. A second reason, which Bladon called *reduction*, is that formants give an incomplete spectral description. Formants are well matched to the spectral peaks, but do not adequately characterize the space between peaks, which can sometimes affect vowel quality. This may explain why, for example, in human listening experiments vowels synthesized from the Peterson and Barney formant data have consistently lower identification scores than in listening tests with the original speech stimuli. In addition, vowels with identical F1 and F2 values, but which are identified as different vowels, can be constructed. This phenomenon is closely related to the spectral center of gravity effect and is discussed in Section 3.5.1. Formants, therefore, do not provide a complete description of the vowel spectrum. A third reason for preferring gross spectral shape, *determinacy*, is that formants are difficult to determine in many situations. This fact will be discussed in more detail in Section

3.2.1. It will be seen that determination of gross spectral shape is often much easier than determination of formants. Finally, *perceptual adequacy*, the ability of a theory to predict perceptual distances between vowels, is greater for gross spectral shape than for formants.

The spectral center of gravity effect, in which nearby spectral peaks are averaged into a single peak for the purpose of vowel recognition, provides another striking example of how gross spectral shape may be more appropriate than formants for vowel perception and production. In experiments on the spectral center of gravity effect (Chistovich & Lublinskaya, 1979; Chistovich, 1985; Chistovich, Sheikin, & Lublinskaya, 1979; Syrdal & Gopal, 1986), subjects matched a single formant stimulus to a 2-formant stimulus by adjusting the second formant. For closely spaced formants ($<$ 3.5 Bark), both formant amplitudes and frequencies affected vowel quality, consistent with averaging of the peak frequencies. This effect is considered in Section 3.5.

The preceding paragraphs state some very obvious, but important, general considerations of the relationship between representations of vowel spectra based on formants and those based on gross spectral shape. In one sense, formants also give a representation of the gross spectral shape. Formants are usually computed by finding the peaks of an LPC spectrum, and the LPC spectrum is, itself, an approximation of the gross shape of the acoustic spectrum of the speech sound. A complete representation of the LPC spectrum requires a specification of the LP coefficients in order for the LPC spectrum to be recovered unambiguously, and the LP coefficients are not identical with the associated formants. Therefore, this dissertation assumes that the formants (locations of the peaks of the LPC spectrum) are distinct from the LPC spectrum itself.

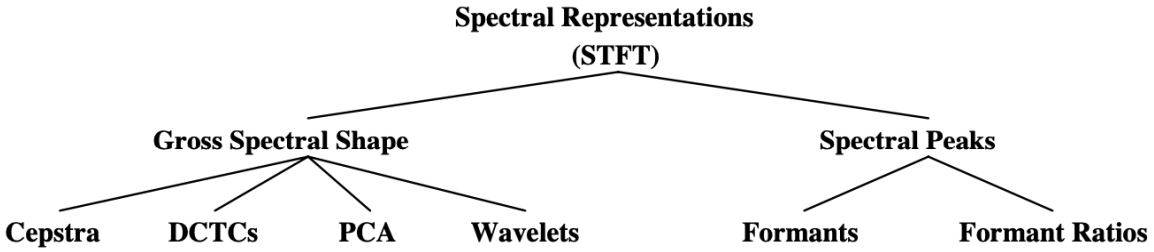


Figure 3·1: Hierarchy of vowel spectral representations. Spectral representations can be divided roughly into those based on gross spectral shape (shown on the left), which include cepstral coefficients, discrete cosine transform coefficients (DCTCs), principal components analysis (PCA), and wavelets, and the locations of spectral peaks (shown on the right) which include formants and formant ratios. The proposed wavelet auditory representation of vowel spectra is an example of a wavelet representation of gross spectral shape.

The next section builds upon these observations by considering, in detail, several representations of vowel spectra.

3.2 Survey of Vowel Spectral Representations

A number of representations of the vowel spectrum have been proposed and explored by speech researchers, each having its own advantages and disadvantages. This thesis assumes that the planning space for speech production is derived from the corresponding perceptual space, making a study of vowel spectral representations crucial for understanding speech production. Figure 3·1 illustrates the rough division of spectral representations into those based on gross spectral shape (on the left) and location of spectral peaks (on the right). Examples of representations based on gross spectral shape include discrete cosine transform coefficients (DCTCs), cepstral coefficients, principal components analysis (PCA), and wavelet-based models (including the wavelet auditory representation proposed in this dissertation). Examples of representations based on location of spectra peaks include formants and formant ratios. The short-time Fourier transform (STFT) is at the top of the hierarchy because it

contains all information necessary for extracting both the gross spectral shape and locations of spectral peaks.

This section presents details of several of these important vowel spectral representations. First the formant and formant ratio representations are examined. Then representations based on gross spectral shape, e.g., principal components analysis and the discrete cosine transform coefficients, are considered. The advantages and disadvantages of each are discussed.

3.2.1 Formant Representation of Vowel Spectra

To a first approximation, a vowel sound is characterized by a steady state Fourier magnitude spectrum. But the unsmoothed spectrum contains more information (e.g., pitch periods) than is necessary for vowel perception. The acoustic theory of speech production (Fant, 1970) models the vocal tract with a rational transfer function whose poles correspond to the formants, and leads to a smoothed approximation of the Fourier magnitude spectrum. The formants are commonly extracted from a speech signal by computing an LPC model (Linear Predictive Code) for the spectrum and then applying a peak-finding algorithm on the resulting LPC spectrum (Rabiner & Schafer, 1978). While automatic formant extraction is typically error prone, requiring much human intervention, with sufficient effort good formant computation can be achieved (e.g., Zahorian & Jagharghi, 1993).

Computing formants is difficult because formant tracking is required. The formant computation algorithm used by Zahorian and Jagharghi (1993) is the following: Speech is digitally low-pass filtered at 3.8kHz and resampled at 8kHz, and then it is high-frequency preemphasized. The speech signal is then windowed with a 50ms Hamming window and a tenth-order LPC (linear predictive coding) model is computed. The roots of the LP polynomial are computed to determine up to five formant

candidates. Actual formants are selected using dynamic programming which "selects the lowest-cost path among the set of formant candidates over the tracking interval" (page 1968) using local costs and transition costs. The algorithm is not fully automatic and performance of the formant tracker must be verified by visual inspection.

Although formants do not uniquely identify vowels across all speakers, they are often used to identify vowels within groups of speakers having approximately the same vocal tract length (e.g., within males, females, and children). Therefore, it is natural to consider formants as a basis for speech production planning within these groups. Simulation results for a formant planning space were presented in Chapter 2 and have been reported elsewhere (Guenther & Johnson, 1995; Johnson & Guenther, 1995) and demonstrate that motor-equivalent vowel production can be achieved with a simple acoustic-like planning space. In that work, a formant planning space based on $F1$ and $F2$ was used to drive a seven degree-of-freedom articulator system (Maeda, 1990) to produce vowels. Difference vectors in the formant planning space were mapped through a learned inverse kinematic mapping to obtain the trajectory in articulator space. The resulting articulator configurations were then converted back to formants through an independent lookup table, and these were used to drive a formant synthesizer (Klatt, 1980).

Although a formant space can be used successfully to produce vowels, a planning space based on a speaker-independent representation of vowels (e.g., using formant ratios) is more desirable. To see why this is true, consider the babbling process used by the DIVA model (Guenther, 1995b), where it is assumed that *adult* speech sounds cause the perceptual system to largely self-organize prior to the onset of babbling. During babbling, however, the speech sounds of the *infant* are used to self-organize the neural mappings needed for movement control in the model. Since both adult

and infant speech sounds participate in the development process, speaker-independent perceptual and planning frames are required. The next section considers a planning space which is more nearly speaker-independent.

3.2.2 Formant Ratio Representation

It has long been known that it is not absolute formants, but formant structure (i.e., the relationship between formants of a vowel) that conveys the linguistic information contained in vowels (Ladefoged & Broadbent, 1957). In fact, a formant ratio theory was proposed as early as the late 1800s. (For a brief history, see Miller, 1989.) It is now also known that speaker pitch (the fundamental frequency $F0$ of the glottal source) plays an important role in vowel perception. Miller (1989) has developed a theory of speaker-independent vowel perception based on formant and pitch ratios, or, equivalently, differences in these values on a log frequency scale. Miller showed that formant ratios corresponding to a given vowel are nearly identical for males, females, and children, even though speakers from these classes have vocal tracts of different lengths. It is interesting that auditory cortex represents sounds using a tonotopic axis that maps to a similar log frequency scale (Shamma, Fleshman, Wiser, & Versnel, 1993).

Guenther et al. (1998) have developed a planning space for vowel production based on Miller's formant ratio model of vowel perception. One justification for considering formant ratios is that these ratios are nearly constant across a broad range of speakers with vocal tracts of different lengths. The model of Guenther et al. (1998) achieves motor equivalence results comparable to the pure formant space, as well as other results. It uses much the same simulation environment as that used in the earlier absolute formant study, with two notable exceptions. First, a radial basis function (RBF) network is used to implement the inverse kinematic mapping. This

eliminates the need for a tessellated map and significantly simplifies the babbling process. Second, another RBF network is used in a forward map to convert from articulator configuration to formant ratios. While the RBF networks reduce the speed of the simulations, the mappings self-organize and reduce the overall complexity of the simulation software.

Miller's perceptual model employs three variables:

$$x = \log(\mathrm{SF3}/\mathrm{SF2}) \tag{3.1}$$

$$y = \log(\mathrm{SF1}/\mathrm{SR}) \tag{3.2}$$

$$z = \log(\mathrm{SF2}/\mathrm{SF1}), \tag{3.3}$$

where

$$\mathrm{SR} = 168(\mathrm{GMF0}/168)^{1/3}, \tag{3.4}$$

and where SF1, SF2, and SF3 are the formants of the vowel, and GMF0 is the geometric mean of the current speaker's voice pitch.

However, Miller (1989) admits that there are problems with his model of vowel perception. One problem is that it does not account for the spectral center of gravity effect (Chistovich & Lublinskaya, 1979; Syrdal & Gopal, 1986) in which nearby peaks tend to be averaged together into one peak for vowel perception. Such a finding suggests that it is gross spectral shape, and not just the location of peaks, that is important for the perception of vowels (see also Zahorian & Jagharghi, 1993). The spectral center of gravity effect is treated further in Section 3.5.

Related to Miller's theory is that of Traunmüller (1984), who proposed that the phonetic quality of vowels is determined by the tonotopic distances between adjacent spectral peaks (e.g., F3-F2, F2-F1, and F1-F0). This hypothesis was corroborated by Fahey, Diehl, and Traunmüller (1996). Although recent papers in the literature

tend to treat F0 as a formant (spectral peak), it cannot be extracted from the speech signal by an LPC analysis in the same way as the other formants, and F0 does not correspond to a vocal tract resonance.

3.2.3 Fourier Transform Spectrum

The short-time Fourier transform (STFT) of the speech signal is the discrete Fourier transform (DFT) of a time-windowed segment of the time-domain speech signal. Usually only the log magnitude of the STFT is used for studying vowel perception and production.

The STFT is high resolution, usually containing 256 or 512 frequency components. In contrast, the LPC-based representations (formants and formant ratios) are smoothed versions of the DFT spectrum with better resolution for peaks, and the discrete cosine transform coefficients (described below) are smoothed versions of the DFT spectrum with equal resolution for both peaks and valleys and with better resolution at low frequencies due to bark warping.

3.2.4 Principal Components Analysis of Spectrum

Principal components analysis (PCA) of the speech spectrum yields statistically-independent spectral shape factors (Pols, van der Kamp, & Plomp, 1969; Plomp, Pols, & van de Geer, 1967; Zahorian & Rothenberg, 1981) and constitutes a specification of gross spectral shape for speech. Low-dimensional PCA representations of spectral shape are useful because smoothed vowel spectra are sufficient for predicting vowel identity. Were this not true, the principal components of spectral shape probably would not capture enough variance of the vowel spectrum to be useful for vowel recognition. For example, two-component PCA plots look like Peterson-Barney plots, and adequately segregate the vowel spectra for the purpose of speech recognition.

Recognition scores using PCA decompositions of the vowel spectrum are as good as those using formants.

3.2.5 Discrete Cosine Transform Coefficients

Zahorian has long maintained that gross spectral shape is better correlated to vowel perception than formants (e.g., Zahorian & Rothenberg, 1981). Zahorian and Jagharghi (1993) recently studied a discrete cosine transform coefficients (DCTCs) parameterization of vowel spectra which uses nonlocal basis functions, in a mathematical formalism similar to cepstral coefficients, commonly used in speech recognition (Rabiner & Juang, 1993). The motivation of Zahorian and Jagharghi (1993) for considering this representation is that it captures gross spectral shape better than other representations.

Zahorian and Jagharghi (1993) compared two sets of spectral features, DCTCs of nonlinearly scaled vowel spectra, and spectral peaks (formants). They write: "We hypothesized that overall global spectral shape provides a more complete spectral description than do three formants and therefore classification based on spectral-shape features should be superior to that based on three formants."

Computation of the DCTCs is described by Nossair and Zahorian (1991). Let $H(f)$ denote the magnitude spectrum of a speech frame, $H'(f)$ a nonlinearly amplitude scaled version of $H(f)$, $H'(f')$ a nonlinearly warped version of $H'(f)$, and let $[H'(f')]$ be a portion of $H'(f')$ over a selected frequency range. The DCT coefficients are defined as the a_n's in the equation:

$$[H'(f')] = \sum_{n=1}^{n=N} a_n \cos[(n-1)\pi f'].$$

Log amplitude scaling is used along with bilinear frequency warping, with

$$f' = f + \frac{1}{\pi}\tan^{-1}\left[\frac{0.5\sin(2\pi f)}{1 - 0.5\cos(2\pi f)}\right]$$

The DCTCs can be interpreted in the following manner. DCTC1 describes the constant term (i.e., the DC or average level of the spectrum), DCTC2 gives a measure of spectral tilt, and DCTC3 gives a measure of spectral compactness. Higher order terms provide additional spectral resolution.

One possible criticism of DCTCs is that more than 10 DCTCs are required to achieve the same vowel recognition results that are obtained with only 3 formants. If minimizing the dimensionality of the spectral representation is of importance, then formants might be the preferred representation. However, if increasing the recognition scores is of primary importance, then DCTCs might be a better choice. This is true because adding formants does not improve the recognition scores, whereas adding DCTCs does. In particular, higher formants correspond to higher frequencies in the spectrum, and vowel recognition doesn't significantly benefit by adding more information at higher frequencies of the spectrum. On the other hand, because the DCTCs correspond to mutually orthogonal basis functions, adding more of them will improve the fit between the DCTC representation and the spectral shape, even at the lower spectral frequencies where vowel recognition can benefit. Therefore, if a higher dimensional representation space is permitted, then a space based on DCTCs is better. It will be seen in Section 3.3 that the representation of vowel spectra by primary auditory cortex is high-dimensional, providing further evidence that the brain does not utilize a formant-based representation of vowel spectra. DCTCs, however, are not an attractive candidate for a representation of vowel spectra in the brain because DCTCs use cosine basis functions which have infinite extent.

Summarizing, in comparison with listening results, DCTC results are more highly correlated with listener responses than formant results. Therefore, research with the discrete cosine transform representation of vowel spectra has shown that both global-shape features and formants are adequate parameters for vowel recognition. Formants offer lower dimensionality, but DCTCs ultimately offer higher recognition scores, probably because exact spectral peaks are not required for perception and approximate formants can be computed from DCTCs. The DCTCs multiply cosine basis functions that have infinite extent. However, the brain is unable to utilize basis functions with infinite extent, making bases that have compact support (i.e., that are zero outside of a finite interval) attractive for models of human speech. The next section examines the biological basis for choosing a representation of vowel spectra for speech perception and production.

3.3 Overview of the Auditory System

Although little is currently known about the neurophysiology of speech movement planning (e.g., see Dronkers, 1996, for a short review), much is known about the representation of vowel and consonant sounds by the auditory system. It is hoped that some light can be shed on the design of the speech planning system in the brain by considering constraints imposed by the representation of speech sounds in the auditory system. This section discusses the representation of speech sounds by the auditory system.

For band-limited signals such as speech, the collection of all Fourier log magnitude spectra constitutes a subspace of the space of square-integrable functions, $L^2(R)$ (Wang & Shamma, 1994a), and an arbitrary function in $L^2(R)$ can be represented by a linear combination of appropriately chosen wavelets (Daubechies, 1992). Therefore,

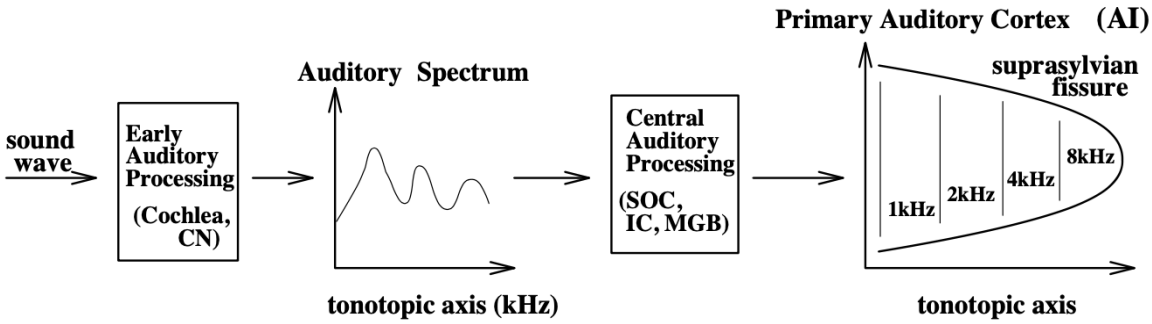


Figure 3·2: Overview of auditory processing. The output of early auditory processing is a representation of the log magnitude short-time Fourier transform of the input sound (Yang, Wang, & Shamma, 1992). This is encoded in primary auditory cortex by a wavelet-like transform (Wang & Shamma, 1994, 1995). The proposed model assumes that speech production targets, and trajectories toward them, are coded in an auditory planning space similar in form to that found in primary auditory cortex.

it is natural to investigate a representation of vowel spectra based on the wavelet transform. The design of the wavelet auditory representation, proposed in this dissertation, is based largely on the physiological modeling work on primary auditory cortex and the auditory periphery by Shihab Shamma and his colleagues.

Consider the simplified block diagram of the auditory system presented in Figure 3·2. Yang et al. (1992) and Wang and Shamma (1994b) discuss processing of sound in the peripheral auditory system and the nature of the signal which is available for encoding in auditory cortex. Many studies of the representation of speech by the early auditory system have been conducted (Sachs & Young, 1979; Ghitza, 1988; Deng, Geisler, & Greenberg, 1988; Cohen, 1989). Yang et al. (1992) present a model of early auditory processing based on the wavelet transform. The output of their model is a representation of the original log magnitude short-time Fourier transform of the stimulus (i.e., all phase information is lost). The correspondence between the output of their model and an input spectrum is difficult to see "because of the complexity of the intervening transformations" (page 827). But this correspondence is established

by the fact that reconstructions from the model output are close replicas of the stimulus spectra, and also by experimental results with automatic speech recognition systems which show that the representation preserves all spectral information and "may even highlight more perceptually useful features" (page 827).

The peripheral auditory system does not compute a perfectly linear Fourier transform of the acoustic signal (Yang et al., 1992). In particular, the frequency axis is logarithmically dilated due to the transduction properties of the basilar membrane in the cochlea. At the cochlear nucleus, the valleys in the spectrum are more depressed, implying an enhancement of the peaks. A cursory inspection of examples of reconstructions from their data (e.g., Figure 7 in Yang, et al., 1992) suggests that this contrast enhancement is moderate, and that using the unenhanced version of the amplitude spectrum (but using the frequency dilation) should give nearly indistinguishable results. Thus, for the purposes of this dissertation it is assumed that it is sufficient to compute the log magnitude of the short-time Fourier transform to preprocess speech for input to the cortical representation.

From their neurophysiological experiments, Shamma and colleagues conclude that at the primary auditory cortex, the sound signal is further coded as a wavelet expansion (Wang & Shamma, 1995, 1994a; Shamma et al., 1993; Shamma, Versnel, & Kowalski, 1995; Shamma & Versnel, 1995). Shamma has demonstrated the existence of cells in primary auditory cortex that have wavelet-like receptive fields (i.e., are self-similar in shape) and suggests that gross spectral shape may play a significant role in perception of vowels and fricatives (Shamma, 1988). In addition, the representation in AI is probably redundant in order to withstand cell death and maintain robustness in noisy environments.

An important feature of the auditory system is the expansion of the 1-D cochlear

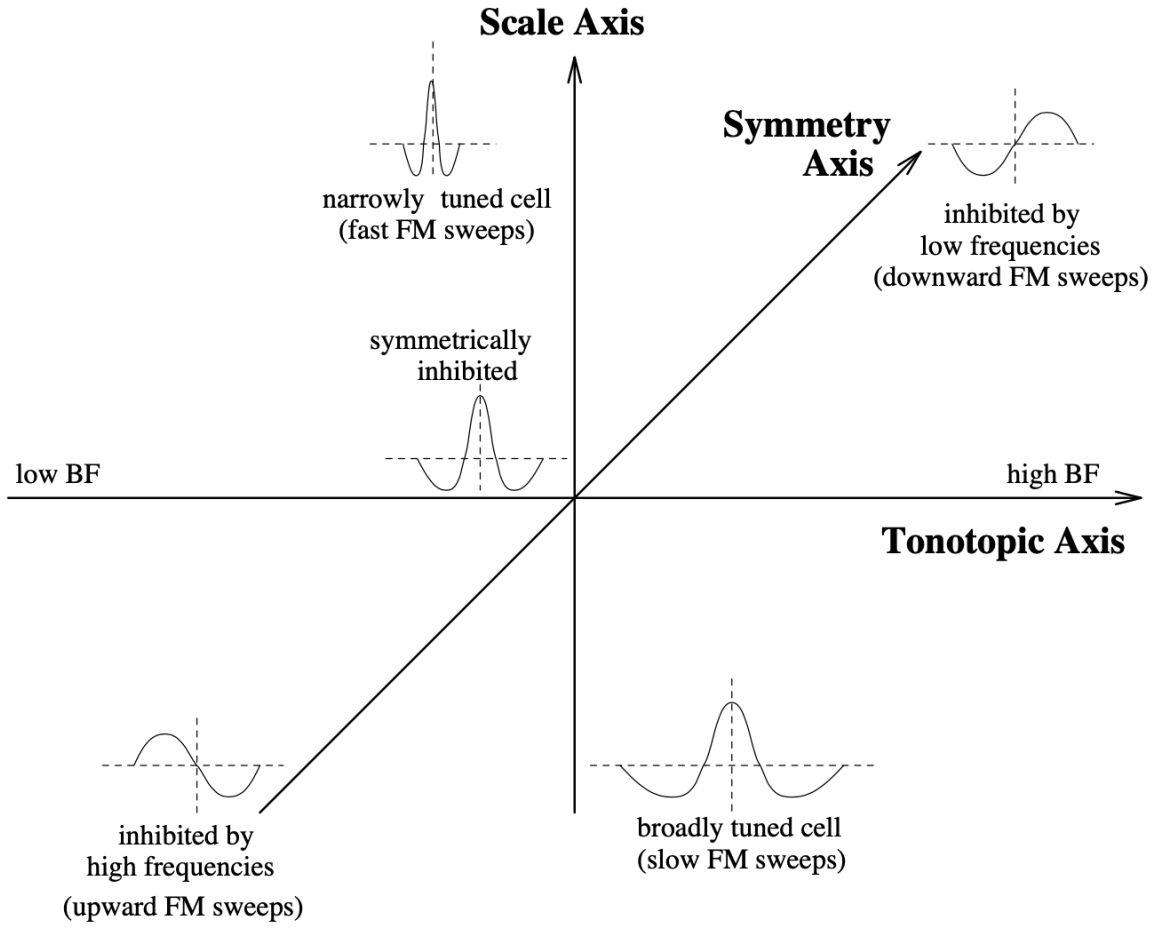


Figure 3·3: Schematic diagram of the three representational axes thought to exist in AI: the tonotopic BF (best frequency) axis, the scale (or bandwidth) axis, and the symmetry axis. (Reproduced from Shamma, 1995, Figure 3.) Physiological experiments reveal that primary auditory cortex (AI) is populated by cells whose responses are selective to a range of spectro-temporal parameters such as spectral bandwidth and asymmetry of spectral peaks, and their transition rates, suitable for extracting and representing spectro-temporal information at various degrees of spectral and temporal resolution.

representation into a 2-D sheet, in which each frequency is represented by an entire sheet of cells (see Figure 3·2). Shamma has studied the functional significance of this expansion (Shamma, 1995) and has proposed the *ripple analysis model*, illustrated in Figure 3·3. Most cells (approximately 90%) in primary auditory cortex are sensitive to local features of the spectrum of the sound stimulus and have a wavelet-like receptive field shape (Shamma et al., 1995). Wang and Shamma (1994a) use Mexican hat wavelets in their modeling of primary auditory cortex. These cells respond best when the sound stimulus has a spectrum with the same shape as the cell's receptive field and is centered at the cell's best frequency (BF). The BF of these auditory cortical cells varies approximately logarithmically across the tonotopic axis. A second axis, perpendicular to the tonotopic axis, codes receptive field bandwidth, or characteristic ripple frequency. Characteristic ripple frequency, Ω_0, varies from 0.2 to 3 cycles/octave, with the average of the distribution being approximately 1.0.

The third axis in the ripple analysis model is the symmetry axis. Cells along this axis vary in the symmetry of their receptive field shape, from symmetrical at the center of AI to asymmetrical at the far ends of the axis (farthest from the center of AI). A mathematical definition of symmetry for AI cells is given in Wang and Shamma (1994a). The symmetry of AI cells has been found to correlate to temporal properties of the spectrum. Shamma et al. (1995) studied the temporal properties of cells in primary auditory cortex of the ferret, and found that temporal features are coded, in part, by basis functions with an asymmetrical spatial receptive field shape. In particular, they considered the response of these asymmetrical cells to swept frequency-modulated (FM) tones. Up and down sweeps were used, and their responses were measured. It was found that cells with greater inhibition at higher frequencies responded better to upward sweeps, and cells with greater inhibition at

lower frequencies responded better to downward sweeps.

Several additional comments on the tonotopic, scale, and symmetry axes are in order. First, Shamma (1995) points out that all three axes are separately capable of coding the shape of a local region of the spectral profile, thereby providing a complementary description of the magnitude of the spectrum. However, the symmetry axis is also capable of coding the phase of the spectrum. Second, formants are relatively broad in bandwidth and thus are represented by the largest scales, i.e., < 2 cycles/octave. Smaller scale cells are capable of coding the fine harmonic structure of the sound. Third, direction sensitivity of AI cells to FM tones has been shown to depend on receptive field asymmetry (Shamma et al., 1995). However, this thesis is not concerned with FM direction sensitivity. Heil, Langner, and Scheich (1992) have shown that sensitivity of AI cells to FM sweep rate is correlated to receptive field bandwidth. Experimental results show that the largest bandwidth cells are also the most sensitive to slowly-varying FM tones, making them ideal for representing the spectra of vowel sounds. Therefore, although the proposed model does not contain an explicit temporal component, temporal features of the cells in primary auditory cortex are important because of their relation to the receptive field bandwidth.

Summarizing, the features of the representation of sound in the auditory system that are most important for the representation of vowel spectra are the use of a wavelet-like representation of the log magnitude Fourier spectrum (using logarithmic frequency scaling), selection of primarily the largest spectral scales in this representation (< 2 cycles/octave), and the use of basis functions with symmetrical receptive field shape. The next section builds upon these observations, and the psychophysical results mentioned earlier concerning gross spectral shape, and proposes a wavelet auditory representation of sound spectra for vowel perception and production.

3.4 Wavelet Auditory Representation of Vowel Spectra

A representation of the vowel spectrum suitable for speech production planning should have the following properties. First, it should capture the right amount of information about the vowel spectrum to enable the speech production apparatus to produce vowels that are correctly recognized by the listener. Psychophysical results discussed earlier suggest that this information is probably the gross spectral shape, not precise locations of spectral peaks. Second, the representation should be easy to compute. Formants, it has been seen, are difficult to determine in many situations, whereas gross spectral shape can be robustly computed by a variety of techniques. Third, the representation should conform to known neurophysiology. Not all representations of gross spectral shape are suitable for a model of human speech production, and the previous section suggested that the representation used by the brain is based on the wavelet transform. This section begins with a discussion of the wavelet transform, then presents the design of the wavelet auditory representation based on an orthogonal wavelet transform, and concludes by considering the effects of an orthogonal versus a redundant representation of the vowel spectrum.

3.4.1 The Discrete Wavelet Transform

A wavelet is any function $\psi(x) \in L^2(R)$ satisfying the "admissibility condition"

$$C_\psi = \int_{-\infty}^{\infty} \frac{|\Psi(\omega)|^2}{|\omega|} d\omega < \infty \tag{3.5}$$

where $\Psi(\omega)$ is the Fourier transform of $\psi(x)$. However, in practice $\Psi(\omega)$ will always have sufficient decay so that the admissibility condition reduces to the following

requirement (Vetterli & Kovačević, 1995):

$$\int_{-\infty}^{\infty} \psi(x)dx = \Psi(0) = 0. \tag{3.6}$$

In other words, a wavelet is a square-integrable function with integral zero.

The wavelet transform of a signal is a weighted sum (or integral) of translates and dilates of a single "mother wavelet". This differs from a short-time Fourier transform (STFT, also known as the Gabor transform) in that the STFT uses a fixed-size window, and the basis functions, therefore, are not constant Q. The STFT suffers from a problem that results from the Heisenberg uncertainty principle, namely that time and frequency resolution must be traded off, and the time (or space) resolution of the STFT is fixed by the use of a fixed-size window. This implies that the frequency resolution is also fixed. For small frequencies, this fixed resolution will often be unacceptable. The advantage of a transform with self-similar basis functions, in which the window size scales with the frequency, is that the frequency resolution can be made proportional to frequency (Rioul & Vetterli, 1991).

The discrete wavelet transform (DWT) is a sampled version of the continuous wavelet transform, an integral transform introduced by Grossmann and Morlet (1984). An N-point DWT can be computed most efficiently by a recursive algorithm of no worse than $O(N \log N)$ derived by Mallat (1989). Daubechies (1988) showed that orthogonal wavelets with compact support exist (i.e., wavelets that are zero outside a finite interval) and provided a technique for finding them.

See Appendix B for a detailed definition of wavelets and the wavelet transform.

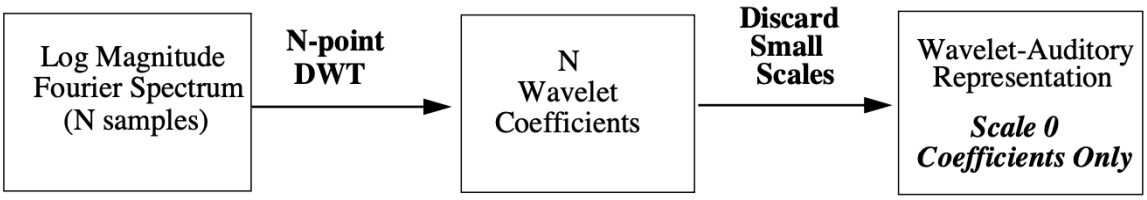


Figure 3·4: Computation of the wavelet auditory representation. An N-point discrete wavelet transform (DWT) is applied to N equally-spaced samples of the log magnitude of the Fourier spectrum of a sound (left block), producing N wavelet coefficients (center block). All but the Scale 0 coefficients are discarded, resulting in the wavelet auditory representation (right block).

3.4.2 Design of the Wavelet Auditory Representation

The wavelet auditory planning space proposed in this thesis is an abstraction that captures some of the essential features thought by some to exist in primary auditory cortex and in subsequent stages of processing within the brain. Problems with formants, and physiological and psychophysical justification for considering a wavelet representation, were presented in Section 3.1. The representation used by the auditory cortex for encoding sounds (including vowel sounds) is described in Section 3.3 and is found to be a redundant wavelet-like representation. We will assume that the wavelet-like representation found in auditory cortex can be approximated by an orthogonal wavelet expansion of the log magnitude Fourier spectrum of the vowel sound.

The wavelet auditory representation is computed by the following algorithm (see Figure 3·4). (1) Begin with N equally-spaced samples of the log magnitude short-time Fourier transform (STFT) of the vowel sound. Two versions of the wavelet auditory representation, corresponding to $N = 128$ and $N = 256$, are examined in this dissertation. The separation between samples is constant between 0 and 4000 Hz (as measured in Bark units). (2) Compute the N-point discrete wavelet transform

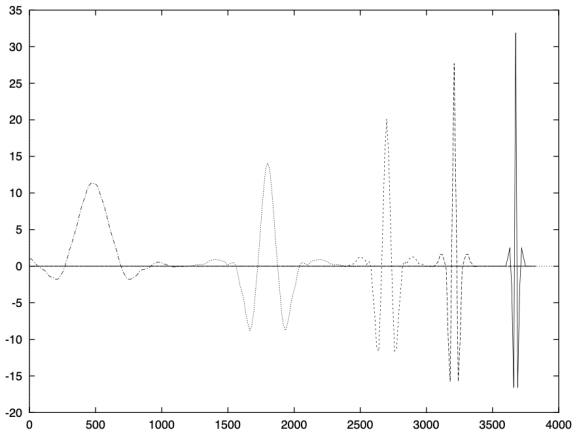


Figure 3·5: Example basis functions used in the definition of the wavelet auditory representation. These basis functions and their translates span the space of log magnitude Fourier spectra. The basis functions used in this model are orthonormal.

(DWT) using the recursive algorithm described by Mallat (1989) using computer software available from Stan Sclaroff (Pentland, 1994). This software produces N wavelet coefficients. A 128-point DWT with 4 wavelet levels consists of 8 scaling functions (see Appendix B for the definition of scaling function) at Scale 0, 8 wavelets at Scale 1, 16 wavelets at Scale 2, 32 wavelets at Scale 3, and 64 wavelets at Scale 4, for a total of 128 basis functions. (3) Discard all but the Scale 0 coefficients. The remaining (Scale 0) coefficients constitute the wavelet auditory representation.

The above DWT employs wavelet basis functions with symmetrical shape, as suggested by the scale axis in Shamma's ripple analysis model. Examples of these wavelets are illustrated in Figure 3·5. Only the largest scale basis functions are used by the proposed model.

These basis functions span a space of possible Fourier log magnitude spectra. An example of one such spectrum, shown in Figure 3·6, represents the spectrum of a steady /A/ vowel sound. In this model, it is assumed that the target of vowel production is a region in spectral space, and that the dimensions of this space are

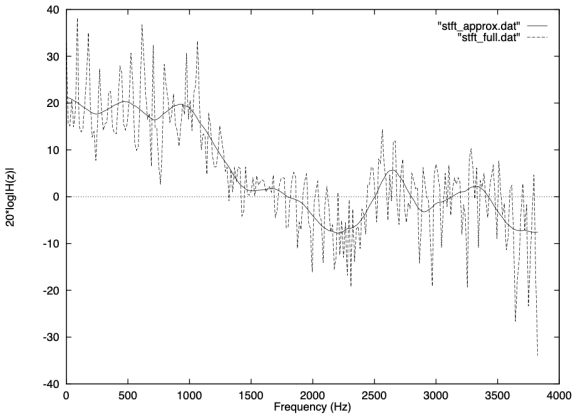


Figure 3·6: Short-time Fourier transform of digitized /A/ sound and its wavelet-smoothed approximation. The dotted line represents a plot of $20 \log |S(e^{j\omega})|$, where $S(e^{j\omega})$ is the Fourier transform of the /A/ sound. The solid line represents a wavelet approximation of the dotted-line function derived from its discrete wavelet transform using only 16 of the 256 possible wavelet basis functions. The wavelet approximation is a smoothed version of the original, and preserves the gross shape of the spectrum.

the wavelet basis function coefficients of the wavelet expansion of the spectrum. For simplicity, it has been assumed that the targets are *points* in this space, but region targets are easily implemented.

In computing the Fourier spectrum, a 512-point DFT is used. Because the time-domain sound signal is real, it is known that the DFT is symmetric about zero frequency. Therefore, only 256 components of the DFT need to be computed and stored. Because the original /A/ sound was sampled at 8000Hz (corresponding to a sampling period of 125μS), the frequency range for the DFT is $0 - 4000$Hz. Therefore, the wavelet basis functions must span this range of frequencies. Under these conditions, perfect reconstruction of the original spectrum is possible using a 256-wavelet DWT.

In Figure 3·6, the dotted line represents perfect reconstruction of the Fourier transform of the /A/ sound using all 256 wavelet basis functions. Using fewer basis functions will result in a smoothed approximation of the original. In the same figure, the solid line represents the smoothed spectrum that results from using the 16 largest

scale wavelets. This will be referred to as the *16-basis wavelet auditory representation* of vowel spectra. If only the largest scales are used, then an approximation which preserves the gross shape of the original spectrum is obtained.

An 8-*basis wavelet auditory representation*, derived from a 128-wavelet DWT, is also evaluated. The motivation for the 8-basis representation comes from the spectral center of gravity effect. In this effect, averaging of spectral peaks occurs over an interval of $3 - 3.5$ Bark (Chistovich & Lublinskaya, 1979; Syrdal & Gopal, 1986). In Section 3.5 it will be shown that the 8-basis representation utilizes wavelets with bandwidth $\approx$ 3.5 Bark and, therefore, may explain the spectral center of gravity effect.

In simulations with this model (described in Chapter 4), vowels of acceptable quality are produced using both the 16-basis wavelet auditory representation, derived from a 256 wavelet DWT, and the 8-basis wavelet auditory representation, derived from a 128 wavelet DWT. An additional constraint on the frequency scaling used in the wavelet auditory representation, namely the Bark scale, is detailed in Section 3.5.

3.4.3 Orthogonal Versus Redundant Wavelet Bases

Not all sets of wavelet basis functions ψ_i yield stable reconstruction of arbitrary functions. Reconstruction of the original function is stable if small perturbations of the wavelet coefficients give rise to small perturbations of the reconstructed function. Sets of basis functions ψ_i for which stable reconstruction is possible are called *frames*. It is easy to show that a set of orthogonal wavelets is automatically a frame, but many sets of redundant (non-orthogonal) bases also constitute a frame (Daubechies, 1992). Satisfying the requirement that a set of basis functions constitute a frame is important in order to ensure stability of the representation.

A set of basis functions is overcomplete or *redundant* if the basis functions ψ_i are

linearly dependent, i.e., if one of the ψ_i can be written as a linear combination of the others. In this case, there are more basis functions than is necessary to span the space of functions. An orthogonal set is minimal because the basis functions are linearly independent. Shamma et al. (1995) claim that the representation in primary auditory cortex is redundant, and therefore, not orthogonal. There are several reasons to expect this result. One is that the loss of a single orthogonal basis function (due to cell death, for example) would make it impossible to span the space of functions previously spanned. Another reason is that the effect of noise at the neuronal level is reduced with an overcomplete set of basis functions. This is suggested by the fact that lower precision for wavelet coefficients is required to obtain the same reconstruction accuracy when using a redundant basis (Daubechies, 1992). Since this dissertation is not concerned with cell death or noise robustness, an orthogonal basis is adequate. What can a model which uses an orthogonal basis reveal about a system that is most likely overcomplete? The sampled values of the wavelet-smoothed spectrum at a given resolution are not affected by the fact that an overcomplete frame is used instead of an orthogonal basis. In other words, a reconstruction of the smoothed spectrum would be the same whether the wavelet transform used an orthogonal or redundant set of basis functions. Therefore, the speech production results presented in this dissertation, which rely on the shape of the spectrum of the speech sounds, should not be affected by the use of an orthogonal basis in the spectral representation.

3.5 A Wavelet-Based Model of the Spectral Center of Gravity Effect

This chapter has motivated and presented the wavelet auditory representation of vowel spectra, and psychophysical and physiological evidence was advanced to argue

for the existence of this representation in the human brain. This representation consists of multiple scales of varying bandwidths and is sufficient to accurately represent the log magnitude of a vowel's short-time Fourier spectrum. In Chapter 4, a version of this representation, using only the largest spectral scales, will be used to plan vocal articulatory movements for the production of vowel sounds. In this section we will study the role of spectral scales in the well-known spectral center of gravity effect.

What is the rationale for using only the largest spectral scales in the wavelet auditory representation? Some of that evidence was presented previously, especially the experimental results of Zahorian and Jagharghi (1993) on vowel perception. More evidence comes from a consideration of the spectral center of gravity effect. The present section examines this evidence and proposes an explanation of the spectral center of gravity effect based on the wavelet auditory representation.

3.5.1 The Spectral Center of Gravity Effect

Two-formant vowel stimuli whose formant peaks are sufficiently close are matched with single-formant stimuli in listening experiments, suggesting that the auditory system performs an averaging or smoothing process on vowel spectra. This phenomenon has been well known since the work of Delattre, Liberman, Cooper, and Gerstman (1952) on vowel synthesis. Chistovich and Lublinskaya (1979) found that the critical distance between formants, below which spectral averaging occurs, is $3 - 3.5$ Bark. The Bark scale is described in Section 3.5.2.

This spectral averaging is known as the *spectral center of gravity effect* (Chistovich & Lublinskaya, 1979; Syrdal & Gopal, 1986). The simplest example of the effect is shown in Figure 3·7. When the amplitudes of sufficiently nearby spectral peaks are equal, the resulting smoothed peak has its maximum, F^*, approximately at the midpoint of the two peaks. Figures 3·7(a) and 3·7(b) illustrate the case of un-

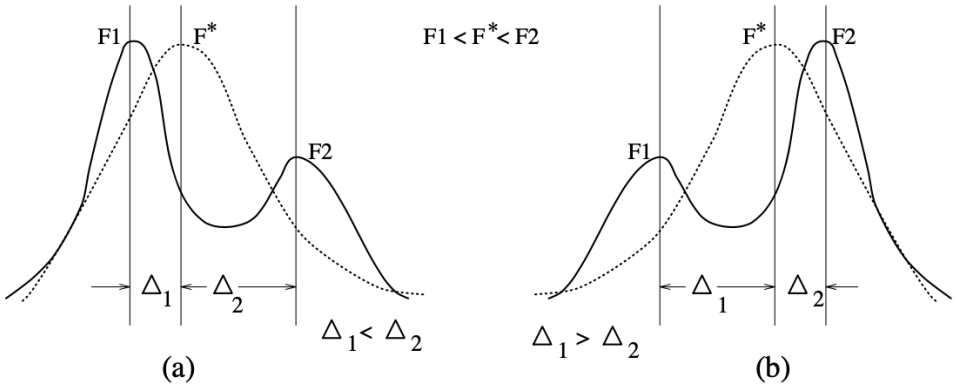


Figure 3·7: Schematization of the spectral center of gravity effect with unequal formant peaks (F1 and F2). Two nearby peaks of the acoustic spectrum (solid line) are smoothed by the auditory system into a single effective peak (dotted line with center frequency F^* such that $F1 < F^* < F2$). By parametrically varying the relative amplitudes of the two peaks, the effective center frequency shifts toward the larger peak. (a) $F1_{\text{amp}} > F2_{\text{amp}}$. (b) $F2_{\text{amp}} > F1_{\text{amp}}$.

equal peak amplitudes, with $F1_{\text{amp}} > F2_{\text{amp}}$ and $F2_{\text{amp}} > F1_{\text{amp}}$ respectively. These cases of equal and unequal amplitude were studied by Bedrov, Chistovich, and Sheikin (1976) and are described in Chistovich and Lublinskaya (1979). There it is demonstrated that the matching single peak can be shifted from one formant to the other, within the $3 - 3.5$ Bark range, by adjusting the ratio of the amplitudes $F1_{\text{amp}}/F2_{\text{amp}}$.

The main contribution of Chistovich and Lublinskaya (1979) is their demonstration of a threshold of $3-3.5$ Bark for the spectral center of gravity effect. Peaks with separation $\Delta z < \Delta z_c$, where Δz_c is the critical distance (≈ 3.5 Bark) are averaged, whereas peaks with separation $\Delta z > \Delta z_c$ are not. Their research concerned Russian vowels, but they argue (and it is assumed here) that their results apply to "quite different languages" (page 194).

Back vowels have formants F1 and F2 separated by less than Δz_c and the relative amplitude of F2 increases from /OO/ to /A/. Therefore, the single-formant approx-

imation of these vowels is possible, with F* being closest to F1 for /OO/ and closest to F2 for /A/. In front vowels, F1 and F2 are separated by more than Δz_c, but F2 and F3 are within Δz_c. Therefore, front vowels have a two-formant approximation, where F2* must be farthest from F2 for /IY/, because $F2_{\text{amp}} < F3_{\text{amp}}$ for /IY/. For /OO/, the best F* $\approx$ 0.3kHz and for /A/ the best F* $\approx$ 1.2kHz (Delattre et al., 1952). All the front vowels can be produced with the same F1 value, while varying only F2* between about 1.4kHz and 3.0kHz (for /IY/).

In addition to the spectral center of gravity effect, any model of vowel perception must also account for the fact that the amplitude of a peak does not affect vowel identity when the peaks are widely separated (and when the peak amplitude is sufficient for detection). As Chistovich and Lublinskaya (1979) write, the data on front and back vowels are "not compatible with the suggestion ... that vowel parameters correspond to some low-frequency components of the whole spectrum shape curve" (page 193).

Chistovich and Lublinskaya (1979) present the outline of a qualitative model which explains the spectral center of gravity effect. Their model consists of a two-stage processor which first extracts the spectral peaks, then averages them over some interval of the frequency axis. This two-stage system is followed by a third stage, a lateral inhibitory network (LIN), which enhances the peaks. By first detecting spectral peaks, their model is able to preserve the peaks even when those peaks are widely separated, and can average peaks together when those peaks are sufficiently close. Qualitatively, the wavelet auditory representation behaves in a similar manner. In particular, the peripheral auditory system is assumed to compute a short-time Fourier transform of the stimulus, and the auditory cortex is assumed to compute a wavelet expansion of this spectrum. The wavelet auditory representation preserves spectral peaks that are

far apart and averages sufficiently nearby peaks into a single peak, thereby conforming to the vowel perception data.

Chistovich and Lublinskaya (1979) suggest that the width of the summation interval, i.e., the interval over which spectral peaks are averaged into a single effective peak, is approximately $3-3.5$ Bark. This implies that the bandwidth of the wavelets in the proposed model must also be about $3-3.5$ Bark. The spectral center of gravity data, therefore, provide an additional constraint on the design of the wavelet auditory representation. Without this constraint, there would be no principled way to choose the bandwidth of the wavelets.

The issue of averaging or smoothing needs to be examined in more detail. Chistovich and Lublinskaya (1979) are correct to worry that only certain kinds of smoothing will explain the spectral center of gravity effect and vowel perception data. In particular, arbitrary low frequency smoothing will not work, in general. For example, the DCTC representation will tend to average nearby spectral peaks, but it will not preserve widely separated peaks as well as the wavelet auditory representation. This is true because the DCTC basis functions are nonlocal (cosines) whereas the wavelet basis functions used in the proposed model are local. In the first stage of the Chistovich and Lublinskaya model, the peaks of the short-time Fourier spectrum are determined, presumably by the standard method of first computing the LPC spectrum (using the appropriate number of LPC coefficients) and then finding the peaks of this spectrum. The process of finding the LPC spectrum is itself a particular kind of smoothing (Rabiner & Schafer, 1978). This is followed by a method they call *summation* which, by implication from their description, is very similar to convolution of the LPC spectrum with a kernel of finite width. Convolution with the wavelet basis functions will also have the desired behavior.

The wavelet smoothing of the vowel spectrum implies a loss of formant resolution which is not observed in nonspeech stimuli (Traunmüller, 1982). However, this is consistent with the view that the proposed wavelet auditory representation of vowel spectra is cortical in nature, and exists in parallel with representations of nonspeech stimuli. It is important to point out that the proposed model is not a complete model of vowel perception. Rather, it is a *representation* of vowel spectra suitable for vowel production, and may exist in parallel with other representations. Indeed, because auditory cortex also codes small spectral scales, it may be true that these smaller scales also participate in the perception of vowels.

3.5.2 The Bark Scale

This section describes the Bark frequency scale, used in all computer simulations discussed in this and subsequent chapters. The Bark scale, presented in detail in Syrdal and Gopal (1986), is based on a functional model of the auditory system as a series of internal overlapping bandpass filters, each of which corresponds to a critical band. Acoustic energy falling within a critical band is averaged. Each critical band corresponds to a fixed length of about 1.3 mm along the basilar membrane, or about 1300 cochlear neurons. The critical band scale increases linearly up to about 500 Hz and then increases logarithmically thereafter. The Bark scale (Zwicker, 1961), named after Barkhausen, the creator of the unit of loudness, is one particular instance of a critical band scale.

In the following simulations, frequencies are converted to Bark according to the Zwicker and Terhardt (1980) approximation, reported in Syrdal and Gopal (1986). The conversion formula is given by:

$$B = 13 \arctan(0.76f) + 3.5 \arctan(f/7.5)^2$$

where f is frequency in kHz and B is critical band value in Bark units. The low frequency end correction is also used; i.e., all frequencies below 150Hz are raised to 150Hz, and frequencies between 150 and 200Hz are corrected according to

$$f_c = f - 0.2(f - 150),$$

and for frequencies between 200 and 250 Hz,

$$f_c = f - 0.2(250 - f),$$

where f_c is the corrected frequency and f is the frequency in Hz.

3.5.3 Computer Simulation Results

The wavelet auditory representation of vowel spectra provides a simple mechanistic explanation for the spectral center of gravity effect. The wavelet transform computes an approximation of the original spectrum at a resolution determined by the scale of the smallest scale wavelet used in the transform. Therefore, if two spectral peaks fall within the bandwidth of this smallest wavelet, these peaks will tend to be averaged into a single peak.

In order to demonstrate how wavelet smoothing of acoustic spectra leads to the spectral center of gravity effect, it is necessary to create spectral stimuli with precise values of the formant frequencies and amplitudes. Although the Maeda articulatory synthesizer (described in Section 2.2.1) is capable of generating examples of spectra with nearby peaks, it does not give enough control over the precise location or amplitude of these peaks. In particular, the Maeda articulatory system does not allow *independent* control over the frequencies and amplitudes of the individual formant peaks, which would be necessary to demonstrate the properties of the spectral center

of gravity effect. Therefore, computer simulations of the spectral center of gravity effect, described in this chapter, utilize a spectrum generator based on Klatt's algorithm (Klatt, 1980). See Appendix A for details of the algorithm. This spectrum generator gives the desired independent control over formant frequencies and amplitudes.

An example of the spectral center of gravity effect, using 3.5 Bark wide basis functions, is illustrated in Figure 3·8. Figure 3·8(a) shows a spectrum with the first two formant peaks differing in frequency by 2.5 Bark and with equal amplitude. Such a spectrum could correspond to a back vowel with relatively high F1 and low F2. When the amplitudes of the adjacent peaks are equal, the resulting smoothed peak has its maximum approximately at the midpoint of the two peaks. Figures 3·8(b) and 3·8(c) illustrate the case of unequal peak amplitudes, with $F1_{\text{amp}} > F2_{\text{amp}}$ and $F2_{\text{amp}} > F1_{\text{amp}}$ respectively. These cases of equal and unequal amplitude were studied by Bedrov et al. (1976) and described in Chistovich and Lublinskaya (1979), and schematized in Figure 3·7. There it was demonstrated that the matching single peak can be shifted from one peak to the other, within the $3 - 3.5$ Bark range, by adjusting the ratio of the amplitudes $F1_{\text{amp}}/F2_{\text{amp}}$. This same parametric behavior is readily demonstrated using the wavelet auditory representation.

The existence of the threshold $\Delta z_c \approx 3.5$ Bark suggests that no smaller scales exist in the representation of vowel spectra, at least for the purpose of vowel classification. In the proposed model, this corresponds to the absence of wavelets with bandwidth smaller than 3.5 Bark. The absence of these smaller scales imposes a limitation on the resolution of vowel spectra that can be represented and presented to the vowel recognition system presumed to exist in the brain. Why are these smaller scales unnecessary for vowel classification? How does the system learn, during development, to ignore these smaller scale spectral features? It is conceivable that some vowel

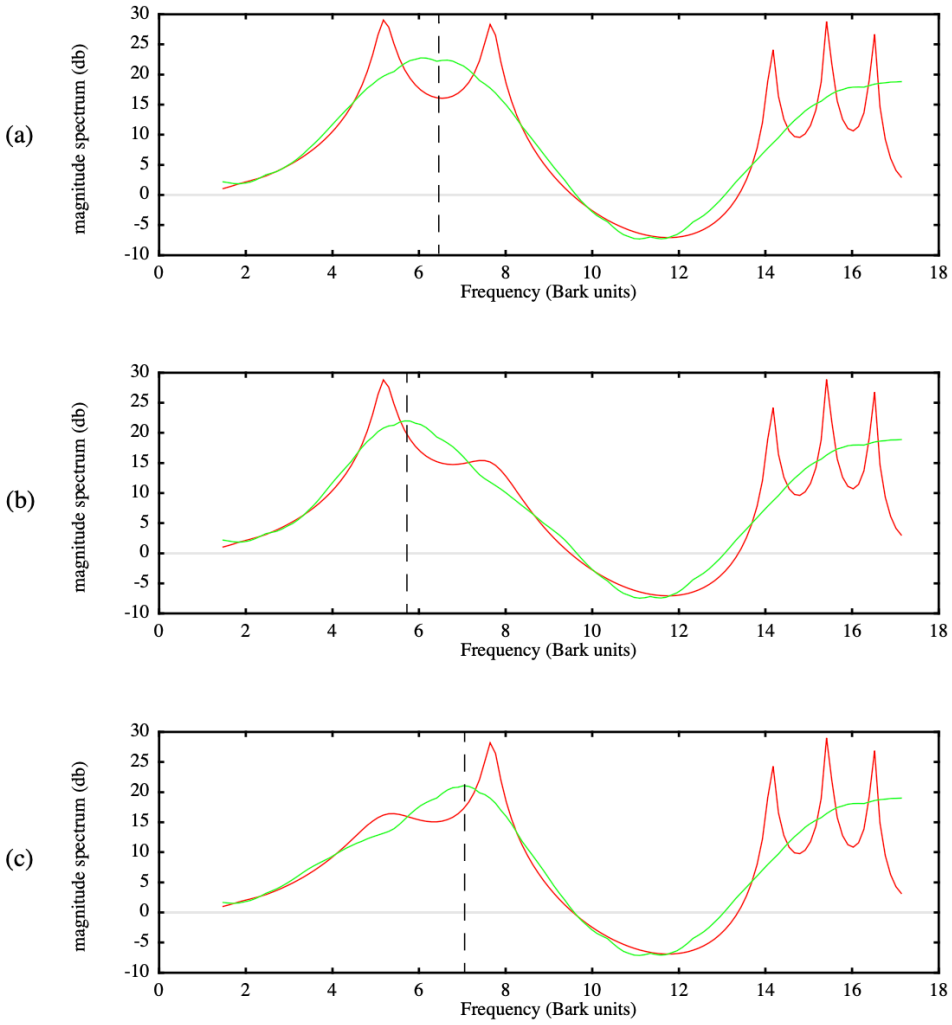


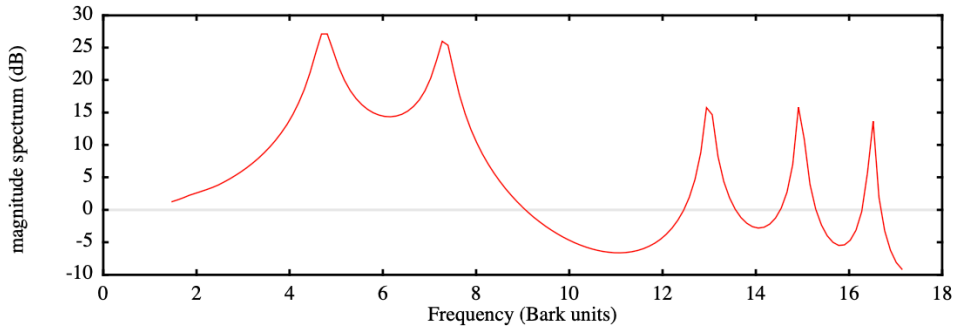
Figure 3·8: Examples of the spectral center of gravity effect. (a) Equal amplitude peaks. (b) $F1_{\text{amp}} > F2_{\text{amp}}$. (c) $F2_{\text{amp}} > F1_{\text{amp}}$. The horizontal axes are in Bark units. In each plot, $F2 - F1 = 2.5$ Bark.

categories will require smaller scales than others. But there will be a "smallest scale" and a "largest scale" applicable to all vowel categories, and this range of scales defines the desired wavelet auditory representation. It is also conceivable that different scales are needed within different frequency ranges.

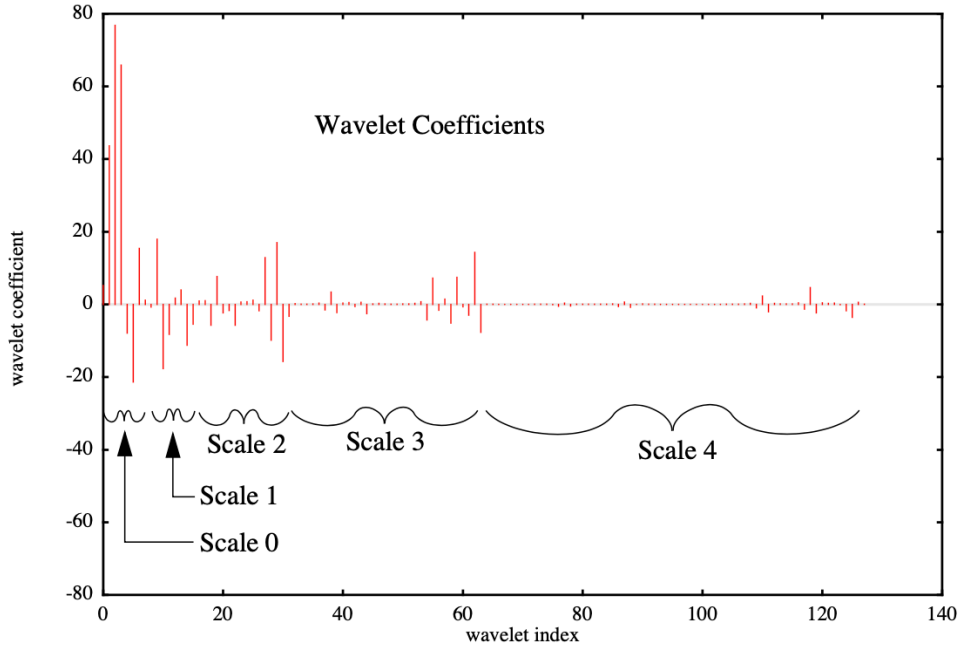
There are two basic reasons why the scales used in the wavelet auditory representation are limited to a narrow range. The first is that typical vowel spectra can be well represented by a wavelet expansion consisting of only these scales. In other words, the coefficients that multiply the wavelets within scales outside this range are small or zero, or, equivalently, the error in the reconstruction of the spectra resulting from omitting these wavelets is small. Physical characteristics of existing human vocal articulators will account for the limited spectral scales seen in vowel spectra of natural speech. Dimensions of the vocal tract, and the frequency range over which speech is analyzed, will impose hard limits on the spectral scales in the vocal tract transfer function.

Figure 3·9 illustrates this effect. Figure 3·9(a) presents a typical spectrum produced by the spectrum generator described in Appendix A and Figure 3·9(b) shows a plot of the wavelet coefficients derived from a 128-point Discrete Wavelet Transform (DWT) of the spectrum in (a). The proposed wavelet auditory representation uses only the 8 basis functions at Scale 0. Wavelets at higher scales would add progressively more detail to the representation. An examination of Figure 3·9(b) reveals that the wavelet coefficients are largest for Scale 0 and are significantly smaller for all other scales. Similar results hold for all spectra considered by the model.

The second reason for a limited range of spectral scales in natural speech concerns the number and distribution of phonemes in spectral space. As long as the few existing phonemes (vowels, in this case) are well separated in the spectral space, using



(a)



(b)

Figure 3·9: (a) A typical vowel-like spectrum. (b) A plot of the corresponding wavelet coefficients. The wavelet coefficients are largest for Scale 0 and are significantly smaller for all other scales, thereby justifying the use of only the largest spectral scale in the wavelet auditory representation.

the given number of spectral scales, then additional spectral scales are unnecessary (Syrdal & Gopal, 1986). Evolution of human speech involves an interplay between the physics of vocal articulators, the number and spectral properties of existing phonemes, and the ability of the speech perceptual system to discriminate these phonemes. It is assumed that these competing factors are in equilibrium, and that the current resolution of vowel spectra, along with the spectral scales that this resolution implies, is adequate. Therefore, while speech researchers may be able to construct spectral stimuli having sufficiently nearby peaks, subjects' vowel recognition systems, which were trained using the natural speech sounds in their native linguistic environment, have no need to distinguish these peaks for the purpose of vowel identification.

3.6 Discussion and Conclusions

The relevance of gross shape of vowel spectra has been known since the time of the vowel synthesis experiments of Delattre et al. (1952), in which nearby spectral peaks can be approximated by a single peak. This phenomenon is known as the spectral center of gravity effect. Bedrov et al. (1976) further explored the spectral center of gravity effect and showed that human listeners perceptually average nearby peaks in vowel recognition experiments, and that the position of this average peak depends on the relative amplitude of nearby peaks. Chistovich and Lublinskaya (1979) confirmed earlier results and showed that the spectral center of gravity effect occurs when the nearby peaks are closer than about 3 Bark units. Syrdal and Gopal (1986) constructed a model of speaker-independent vowel perception using the spectral center of gravity effect as a starting point. These psychophysical results suggest that gross spectral shape is very important for vowel perception.

In addition to the psychophysical results, this chapter mentioned several problems

with formants, including the difficulty of computing formants reliably and the fact that formant peaks are not preserved during a long speech utterance. Zahorian and Jagharghi (1993) examined a DCTC representation of the vowel spectrum and showed that this representation is correlated better than formants with vowel recognition scores, and they argued that gross spectral shape is more important than formants in theories of vowel perception. In addition, practical speech recognition systems use cepstral coefficients, which are based on the gross spectral shape, to code the spectral properties of speech.

Physiological evidence from single cell recordings in primary auditory cortex suggests that the auditory system codes the gross spectral shape of vowel sounds using a wavelet-like expansion of the log magnitude of the short-time Fourier spectrum. The receptive field shape approximates a Mexican hat wavelet having multiple spectral scales.

Based on these psychophysical and physiological data, the wavelet auditory representation of vowel spectra was proposed. The wavelet auditory representation utilizes an orthogonal wavelet transform of the log magnitude short-time Fourier spectrum of a sound (with Bark frequency scaling), and consists of the coefficients of the wavelets having the largest spectral scale. Reconstruction from this representation results in a smoothed approximation having the gross shape of the original spectrum. Two versions of wavelet auditory representation are studied in this dissertation, the first having 16 basis functions and the second having only 8 basis functions. The 8-basis representation uses wavelets with a bandwidth of ≈ 3.5 Bark, and provides a possible explanation of the spectral center of gravity effect.

A wavelet planning space for vowel production offers a number of advantages over formant planning spaces. First, the wavelet-smoothed spectrum preserves gross spec-

tral shape, which is known to correlate better to vowel perception data. The peaks of the wavelet-smoothed spectrum often, but not always, correspond quite closely to the formants of the vowel. Second, a wavelet representation of the log magnitude Fourier spectrum corresponds to known auditory neurophysiology. No such correspondence is known for the formant representation. Third, and perhaps most significant from a practical point of view, the wavelet representation is easier to compute than the formant representation. Accurate formants are difficult to compute and generally require computation of formant trajectories to ensure continuity of the individual formants. This is not necessary with the wavelet representation of the spectrum. Finally, it was demonstrated that the wavelet auditory representation, using the Bark scale, leads naturally to the spectral center of gravity effect. A rationale for the existence of limited spectral scales in representations of vowel spectra, and hence for the existence of the spectral center of gravity effect, was also presented.

Chapter 4

Speech Production Using the Wavelet Auditory Planning Space

The wavelet auditory representation of vowel spectra, described in Chapter 3, was motivated by the physiology of primary auditory cortex, and was used to explain the spectral center of gravity effect. In this chapter, the wavelet auditory representation is used to define a planning space for the production of vowels and stop consonants, and is embedded in the DIVA model. The resulting speech production model shares many features with the formant-based planning model described in Chapter 2, including motor equivalence, compensation with bite blocks, coarticulation effects, and resulting vocal tract shapes similar to humans. Finally, vowel sounds produced by the model are easily identified by human listeners.

4.1 Preliminary Investigations with the Wavelet Planning Space

Before describing the modifications to DIVA required by the wavelet auditory planning space, results from a simpler model will be presented. In this simpler model, an articulatory system and the associated inverse kinematic mapping were not employed. The purpose of experiments with this simpler model was to determine whether linear interpolation in the wavelet auditory space from a starting spectrum to a final spectral target yields vowel transitions and final vowel sounds of acceptable quality.

These experiments also provided an opportunity to gain experience with the discrete wavelet transform (DWT) and the wavelet auditory representation in the context of vowel production.

In these preliminary simulations, a digitized recording of a steady vowel was analyzed to obtain its Fourier transform, and this was converted by the DWT to the corresponding target in the wavelet planning space. A smoothed version of the spectrum was computed from the DWT using only the 16 largest-scale wavelet basis functions, and this spectrum was defined to be the wavelet target for the vowel. This target was instated during the production of the vowel by the system. The starting spectrum was taken to be a flat log magnitude spectrum of 0dB at each frequency. Production of a vowel consisted of computing the PDV (Planning Direction Vector), i.e., the difference in wavelet space between the current wavelet state and the wavelet target. A small fraction of this difference is added to the current spectrum to obtain the new spectrum. In order to produce the output sound, the new wavelet spectrum was converted back to Fourier spectral space by the inverse discrete wavelet transform (IDWT), and that spectrum was inverse transformed back to the time domain by the inverse short-time Fourier transform (ISTFT), assuming zero phase.

During production of the vowel, the spectrum of the output converged rapidly to the target spectrum in the wavelet space. The left side of Figure 4·1 shows a typical /A/ sound in the time domain recorded from the voice of the author, and the right side shows the output of the simplified vowel production system using the 16-basis wavelet auditory planning space. Two factors contribute to the differences between the target time domain signal and the output of the vowel production system. One is that only 16 of the 256 wavelet basis functions are used, resulting in a smoothed version of the original /A/ sound. Another factor is that the phase of the Fourier

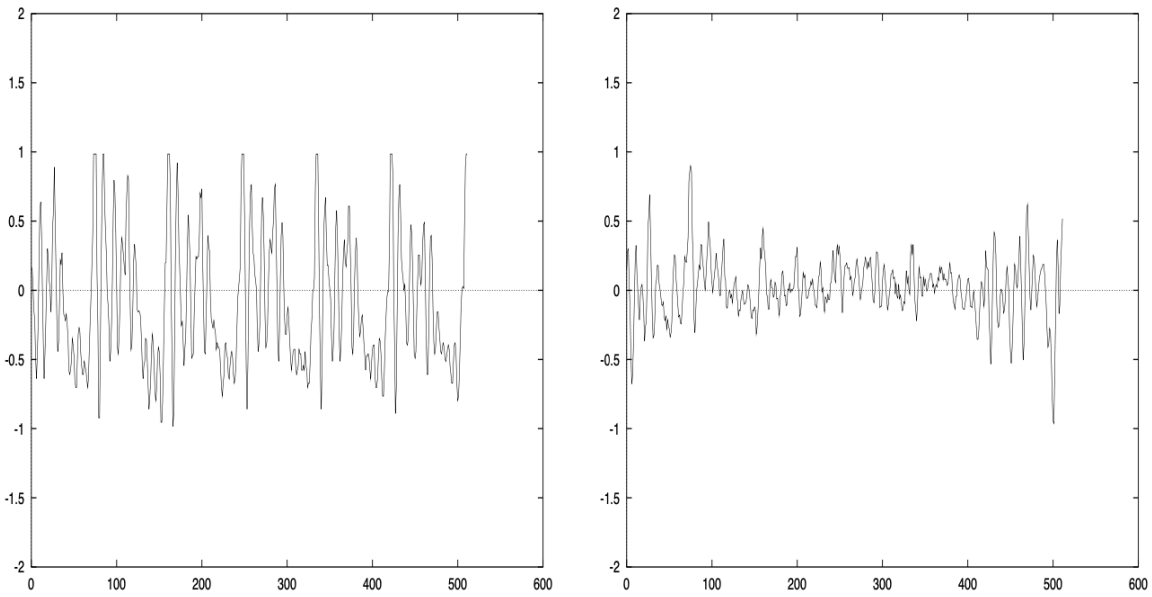


Figure 4·1: The left figure shows 512 time-domain samples of a steady /A/ sound recorded from the author's voice. The corresponding spectrum is computed using the wavelet auditory representation. A straight-line trajectory in the wavelet auditory planning space is planned, from a starting flat spectrum to the target spectrum of the /A/ sound, and intermediate points along the trajectory are converted to speech sounds (see the text for details). The right figure shows the final output of this simplified system. Although there are differences in the final reconstructed time-domain signal from the original, its spectral properties are similar to the original /A/ sound, and informal results suggest that the sound is readily identified as an /A/.

transform is not preserved either in the target or by the production system.

Similar results were obtained for each of 9 English vowels, indicating that linear interpolation in the wavelet auditory planning space would yield acceptable vowel production results. Later simulations with the full DIVA model look more closely at the resulting spectrum for each vowel. The remainder of this chapter is concerned with the DIVA model, modified to use the wavelet auditory representation as the planning reference frame for vowel production.

4.2 Modifications to DIVA Required by a Wavelet Planning Space

The reader unfamiliar with the DIVA model is urged to review Chapter 2 before continuing. Many of the concepts and mathematical equations presented there are relevant for the present version of DIVA, with the exception of issues raised by the following paragraphs.

4.2.1 Articulatory Constraints on Area Functions and Spectra

Linear interpolation within the wavelet auditory planning space, by the method described in the previous section (in which an articulatory system is not used), may result in intermediate spectra that are not physically realizable by the vocal apparatus, even when the start and end points of the linear trajectory correspond to physically realizable vowels. An additional constraint on trajectory formation is required in order to ensure that trajectories in the planning space are physically realizable, and one natural constraint is imposed by the articulatory system itself.

A block diagram of the DIVA model, modified to use the wavelet auditory representation, is given in Figure 4·2. In the DIVA model, a trajectory is formed in the planning space between successive targets by linear interpolation (see also Bullock

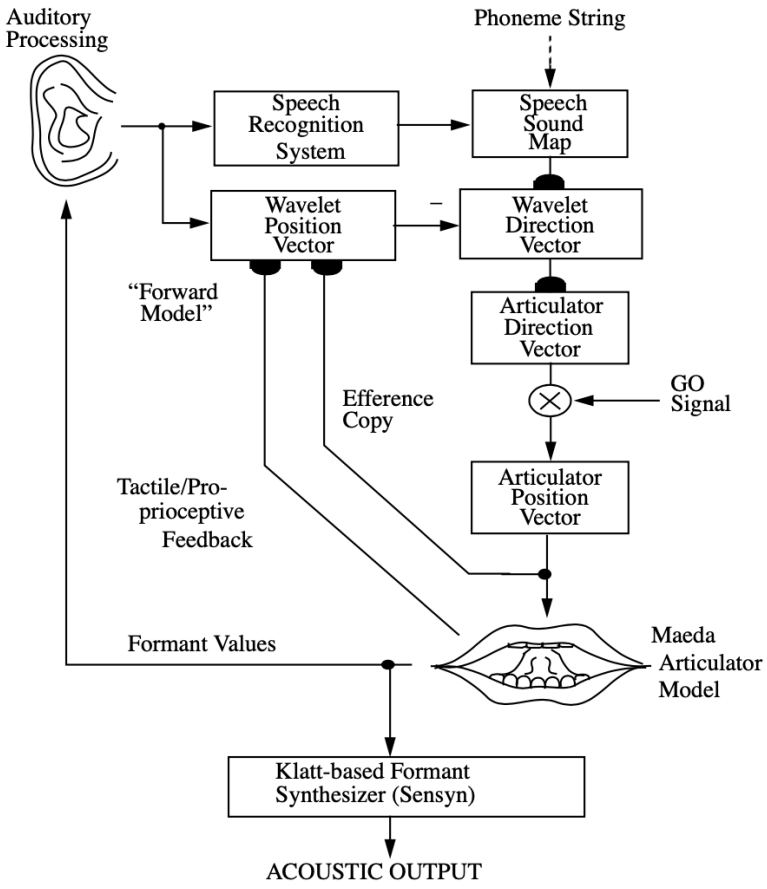


Figure 4·2: Overview of the DIVA model, with modifications to use the wavelet auditory planning space. During the performance phase, desired vowels activate cells in the Speech Sound Map, which read out learned target regions in the planning space. The Planning Direction Vector (PDV), here shown as the Wavelet Direction Vector, is formed by computing the difference between the target region and the Wavelet Position Vector. The PDV is mapped to the corresponding Articulator Direction Vector (ADV) through a learned direction-to-direction mapping, whose weights depend on articulatory position. The ADV is integrated to form the Articulator Position Vector (APV), from which the position in the wavelet auditory space is computed by the Forward Model. A Speech Recognition System identifies the phonemes produced during the babbling phase, whereby random ADVs drive learning of the direction-to-direction mapping and the planning space targets.

& Grossberg, 1988) and mapped through an inverse kinematic mapping to a set of articulators (Guenther, 1995a, 1995b). For speech production, these articulators correspond to the supraglottal vocal tract articulators used to produce vowels and stop consonants.

The primary function of the supraglottal vocal tract articulators is to control vocal tract shape. By the acoustic theory of speech production (Fant, 1970), it is vocal tract shape that is primarily responsible for the static spectral properties of speech. The static acoustic properties of the vocal tract have been successfully explained by treating the vocal tract as a set of concatenated tubes excited by a source waveform (Fant, 1970). This model permits the computation of the spectrum of the resulting speech sound. (See Rabiner and Schafer, 1978, for details of the concatenated tube model.)

However, not all conceivable vocal tract shapes can be realized by the human vocal apparatus. In the case of vowel production, a low-dimensionality model of speech articulation has been derived by Maeda (1990) based on principal components analysis of vocal tract shapes observed during the production of vowel sounds. In his model, the vocal articulators are given by seven articulatory degrees of freedom: (1) jaw height, (2) tongue body position, (3) tongue body shape, (4) tongue tip position, (5) lip aperture, (6) lip protrusion, and (7) larynx height. These articulators account for approximately 90% of the variance seen in vocal tract shape and are closely connected in behavior to the corresponding vocal articulators in the human speech system. The Maeda articulatory system has been described in Section 2.2.1.

A model of speech movement planning must utilize an idealized, but realistic, articulator system in order to ensure that only realistic vocal tract shapes are produced, and, therefore, only realistic vowel and consonant spectra are generated and used in

the feedback to the movement planning system. Not only does the articulatory model provide a way of displaying the output of the model (in the form of articulatory configurations), it also provides input to the model in the form of spectral feedback (via the forward map and via auditory input).

One consequence of using the Maeda articulatory system to compute the spectral feedback in the model is that the spectra produced by the model are likely to be more similar to those obtained from human speech. Another effect is that learning of the forward and inverse maps is more rapid and stable. This is true because a much smaller subspace of spectra contribute to the learning of the maps. It is significantly smaller because it is limited by the possible vocal tract (articulatory) configurations. This means that less exploration of the space is required to learn the map, and there is less interference from other patterns which might disrupt this learning.

Thus, although the planned trajectories are not explicitly constrained to be physically realizable, the actual trajectories produced by the model are constrained by, and limited to, those which can actually be produced by the articulatory system.

4.2.2 Requirements for a Planning Space

Extensive motivation was given for the wavelet auditory representation of vowel spectra in the previous chapter. One of the purposes of the present chapter is to show that the wavelet auditory representation satisfies very general requirements for a speech planning representation. These requirements include:

- Acoustic or auditory representation of the target sounds. This requirement is equivalent to the claim that speech production is planned in an acoustic-like space.

- Low dimensional representation of the planning space. The 8-basis version

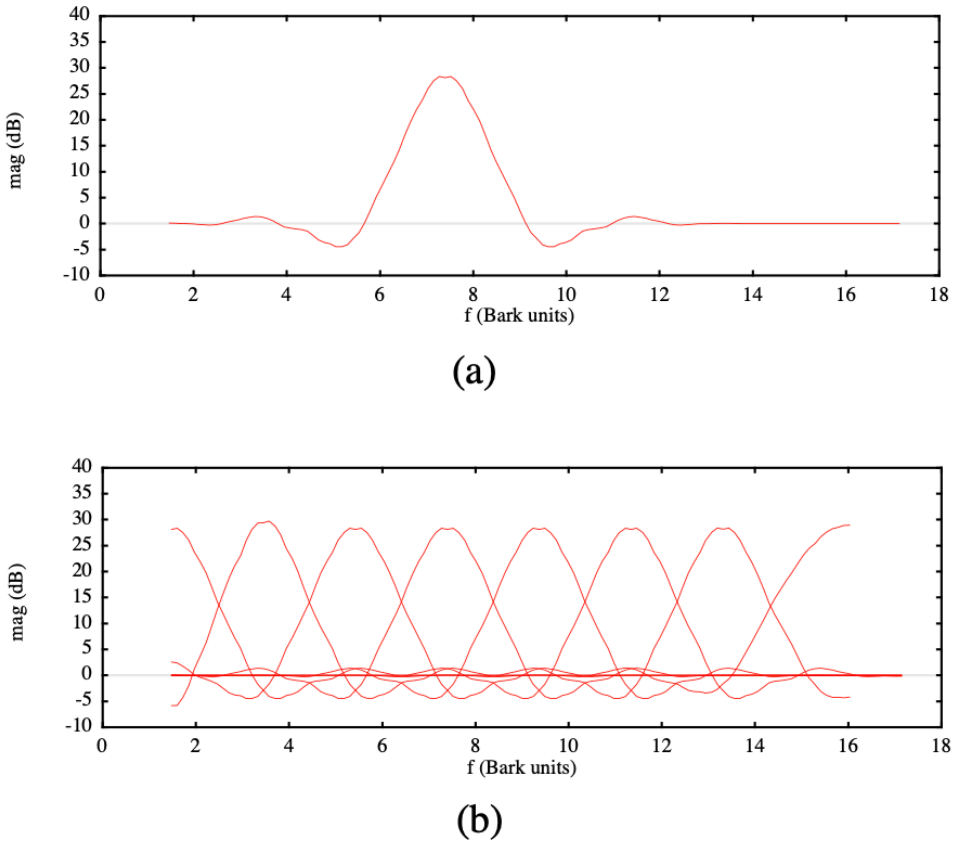


Figure 4·3: Wavelet basis functions used in the model. The model employs a 128-point DWT (Discrete Wavelet Transform) with one level consisting of 8 scaling functions, and four wavelet levels. The present implementation uses only the 8 scaling functions as the planning space dimensions. (a) Single scaling function from the set of 8. (b) 8 orthonormal scaling functions span the space of log magnitude spectra.

of the wavelet auditory representation is used in simulations reported in this chapter. Figure 4·3 illustrates the basis functions used in the wavelet auditory representation. While 8 dimensions is more than the 2 dimensions used in the formant-based model, it is still low enough for efficient and robust production of vowel sounds.

- Low dimensional representation of the inverse kinematic map between the planning space and the articulatory system. Section 4.2.7 will demonstrate that a hyperplane RBF network consisting of only 162 basis functions is sufficient to represent the inverse kinematic map.

- The research presented in this dissertation assumes that the targets of speech movement planning, and the forward and inverse kinematic maps between acoustics and articulation, are not completely hardwired by evolution. Instead, it is assumed that they are tuned during early childhood. Therefore, a teaching signal is required from the perceptual system for learning the forward kinematic mapping from articulatory positions to spectra. This requirement is equivalent to the claim that the form of the planning space follows closely the form of the corresponding perceptual space.

- Representation based on the gross shape of the acoustic spectrum (STFT). This requirement follows from observations from both psychophysics and physiology, and was discussed extensively in the previous chapter.

4.2.3 Spectral Smoothing Implies Uncertainty in Formant Values

A dominant theme of Chapter 3 is that a model of speech perception based on spectral smoothing offers advantages over a formant representation of the speech spectrum, and that the representations used by the vowel production system are derived from the

representations used by the speech perception system. This chapter will demonstrate that the spectral smoothing embodied in the wavelet auditory representation is not in conflict with the requirements of speech movement planning by showing that vowels and stop consonants of acceptable quality can be produced using wavelet movement planning.

However, one feature of formant-based models that must be sacrificed when using the wavelet auditory representation is precise control over formant values. The loss of spectral resolution inherent in the smoothing of the wavelet auditory representation implies uncertainty in the location of spectral peaks. As long as this uncertainty does not prevent production of vowels and consonants, the wavelet auditory representation is viable for speech production.

One consequence of this uncertainty is that similar wavelet spectra may have very different formant values. Indeed, because smoothing may combine nearby peaks into a single spectral peak, the number of formant peaks may not be preserved. However, because of the design of the wavelet auditory representation, which uses basis functions with width ≈3 Bark, it is expected that formant frequencies will not vary by more than a maximum of about 3 Bark units from their expected values, and that formant peaks will be preserved when their spacing exceeds 3 Bark.

4.2.4 Vowel Static Spectral Targets

This and the following sections describe how the wavelet auditory targets are determined for use in the model.

A rectangle in the F1-F2 plane defines the stopping criterion for a given vowel. The stopping criterion is used to determine when the model successfully produces a vowel and causes the utterance to terminate when the values of F1 and F2 fall inside the formant rectangle. Figure 4·4 shows the formant rectangle for each of the

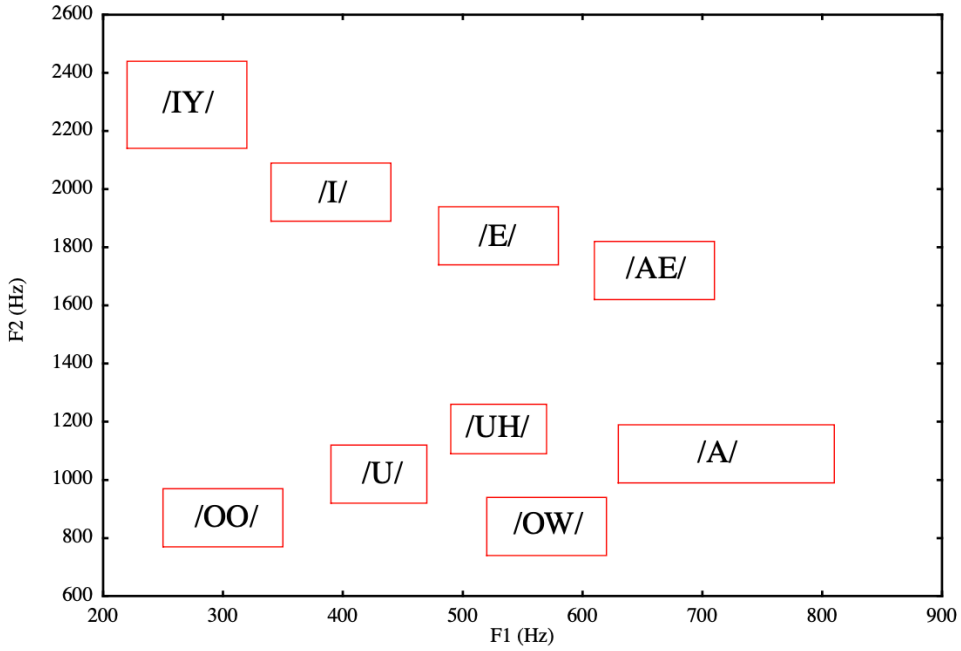


Figure 4·4: Stylized representation of 9 English vowels in formant space. Vowel recognition targets are shown here as rectangles in F1-F2 plane and correspond roughly to the Peterson-Barney data. Vowel production targets are regions in wavelet space. F1-F2 axes not to same scale. The vowels are /IY/ (beet), /I/ (bit), /E/ (bet), /AE/ (bat), /OO/ (boot), /U/ (foot), /UH/ (but), /OW/ (bought), and /A/ (father).

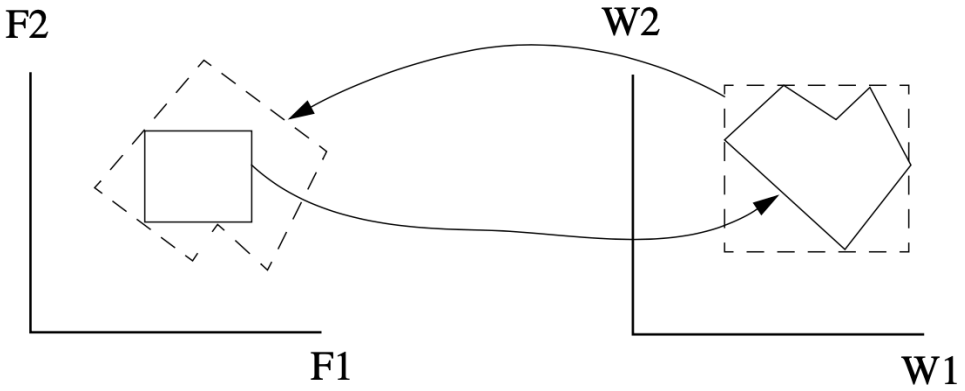


Figure 4·5: Hyper-rectangular wavelet targets derived from formant rectangles are too large. The formant rectangle on the left maps to the polygonal region in wavelet space on the right. By the learning law used by DIVA, this causes the circumscribed hyper-rectangular region in wavelet space to be learned as the target. This is schematized by the rectangle on the right. This region corresponds to the region of formant space schematized by the polygonal region on the left.

9 English vowels produced by the model. The wavelet target of a vowel is defined to be a hyper-rectangular region in wavelet space, and is indirectly derived from the stopping criterion. Therefore, the wavelet target and the stopping criterion are not identical. The wavelet target is determined in the following manner. During a phoneme babbling phase, random movements of the articulators are generated. The stopping criterion rectangle used during babbling is one-half the size of the rectangle used during normal vowel production in order to overcome a problem described in the next paragraph. If the babbled sound has its formants F1 and F2 within this rectangle, the corresponding wavelet state is used to learn the phoneme target by the learning laws described in Chapter 2. During phoneme babbling, each vowel is typically produced (according to the stopping criterion) 10-100 times, and these "hits" cause learning of the phoneme target in wavelet space. More details of the phoneme babbling phase are given in Section 4.2.10.

The set of points in formant space corresponding to a given vowel do not constitute

a rectangular region, but is more appropriately described as elliptical. (For example, see the plot of the Peterson-Barney data in Figure 3.4 of Rabiner and Schafer, 1978.) The DIVA model utilizes hyper-rectangular target regions because of their simplicity, and because very good results are obtained from hyper-rectangular target regions. One possible problem with the approach used in this chapter to learn the wavelet target regions using a formant-based recognizer is that rectangular regions in formant space may lead to hyper-rectangular wavelet target regions that are too large. Consider Figure 4·5. The solid rectangle in the formant plane on the left in Figure 4·5 corresponds to a set of points in wavelet space schematized by the solid *polygon* in the W1-W2 plane. Learning of the target region by DIVA always produces a rectangle (or hyper-rectangle) that circumscribes the points that are sampled during babbling. However, this hyper-rectangle in wavelet space no longer corresponds to the original rectangle in formant space, but maps to the dashed polygon on the left. This dashed polygon is larger than the original formant rectangle, and may cause the DIVA model to terminate vowel utterances before the formant rectangle has been reached. This problem can be avoided by using sufficiently small formant regions during the babbling phase. A better approach would be to define the phoneme targets (for recognition) directly in terms of the wavelet auditory representation. However, this was not necessary because use of small formant target regions during babbling enabled the model to produce all 9 English vowels with and without blocked articulators, while using articulatory configurations typically used by humans.

4.2.5 Stop Consonant Static Spectral Targets

In the present model, the phoneme recognizer utilizes a combination of spectral and vocal tract information, and recognizes both vowels and consonants. Reliable recognition of stop consonants from spectral information alone is a hard problem, in which

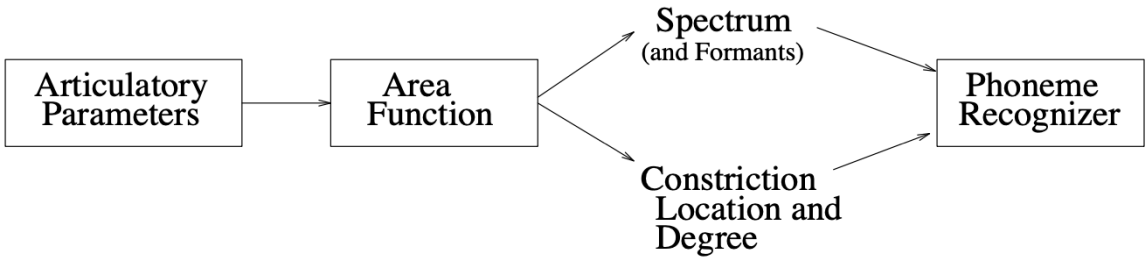


Figure 4·6: Simplified block diagram of the phoneme recognizer used in the model.

many spectral features may be used. Because modeling of stop consonant perception and recognition *per se* are not central to this thesis, and because existence of a phoneme recognizer has been assumed a priori, without loss of generality all necessary spectral *and* articulatory information is used to correctly identify phonemes during babbling and production. In particular, constriction location and degree information is used for the identification of stop consonants.

An approach similar to that used above for vowel sounds is used to learn the static spectral target for each of the three voiced stop consonants, /b/, /d/, and /g/, except that the stopping criterion is defined in terms of the area function (cross sectional area of the vocal tract as a function of position along the vocal tract). A stop consonant is defined to occur when the minimum cross sectional area of the vocal tract is less than a preset small value, and the identity of the stop is determined by the index of the section (tube) along the vocal tract at which this minimum occurs. A simplified block diagram of the phoneme recognizer used in the present model is presented in Figure 4·6.

In order to locate the narrowest constriction of the vocal tract, the area function corresponding to each babbled sound is calculated using a 17-tube model. Table 4.1 is used to define the stopping criterion for each voiced stop consonant. A sound satisfying one of these area function conditions is identified as the corresponding stop

Table 4.1: Area function parameters for voiced stop consonants. (All areas in cm^2.)

Tube #	/b/	/d/	/g/
0-7	> 1.0	> 1.0	> 1.0
8-10	> 1.0	> 1.0	$0.10 < x < 0.21$
11-13	> 1.0	$0.10 < x < 0.21$	> 1.5
14-15	> 1.0	> 0.3	> 1.5
16	$0.07 < x < 0.14$	> 1.0	> 1.5

consonant.

An examination of Table 4.1 reveals that the constriction degree is non-zero for each of the three voiced stops. The choice of the degree of the narrowest constriction for each stop was made primarily to ensure a reasonable spectral target for production. Narrower constrictions have static spectra with very low energy, making it very difficult for the system to distinguish one stop consonant from another. This is true because *any* movement that reduces the total energy of the spectrum will be chosen initially by the model, even when other movements ultimately might be more appropriate for the target stop consonant. For example, if /g/ is desired, the system will initially modify both the tongue body position and lip aperture. If a very low energy spectrum is used for the /g/ target, the lip aperture will continue to decrease until the lips are nearly closed.

This sensitivity of the wavelet auditory representation to total spectral energy might be one disadvantage for stop consonant production (although it is possible to imagine that total spectral energy should be a good indication of constriction degree). However, there is evidence that the early auditory system normalizes the total spectral energy of sounds. Wang and Shamma (1994b) found that the early auditory system "computes a spectrum divided by a smoothed version of itself", which makes "it look less distorted when compared to the original acoustic spectrum" (page 421).

One effect of this normalization is enhancement of the peak-to-valley ratio of the spectrum. Another is relative stability of the auditory representation with respect to overall scaling of the acoustic spectrum. This normalization strategy was not investigated for this dissertation.

This sensitivity to total spectral energy may also be an artifact of the exclusive use of *static* spectral information in the target. Dynamic spectral information (e.g., formant or spectral transitions) are not considered in this dissertation, but may provide additional constraints on the movement planning which could eliminate the problems associated with low spectral energy. Another possibility is that *some* tactile information may participate in movement planning of stop consonants, especially when purely acoustic information is insufficient. Tactile information is not used in this dissertation for movement planning, and the results presented later in this chapter show that static spectral information alone is sufficient to capture *most* static features of vowel and consonant production.

The typical spectral target (for both vowels and consonants) is determined in the following manner. For each babbled sound satisfying a vowel formant stopping criterion, the magnitude of the articulator vector (L^2 norm of the 7 Maeda parameters) is calculated. This magnitude represents the distance of the articulatory configuration from the neutral configuration, and provides one measure of the difficulty of producing that configuration. Within a phoneme, the articulatory configuration having the minimum norm is stored in computer memory from among all hits for the phoneme. This minimum is used to compute the corresponding wavelet auditory spectrum, which then becomes the typical wavelet spectrum (or the spectral target when point targets are used).

The reason for using this minimum length configuration is to ensure that the

spectrum closest to the neutral vocal tract configuration is used. This avoids the learning of targets corresponding to very difficult, or rarely produced, articulatory configurations. An alternative method was also examined, and abandoned, whereby the spectra of all hits for a given phoneme are averaged. However, it was found that this average often corresponds to a spectrum which cannot be realized by the Maeda articulators, even though the spectra contributing to the average are realizable.

4.2.6 The Inverse Kinematic Mapping

The inverse kinematic mapping transforms a desired change in the planning space into the corresponding movements of the articulators needed to achieve this change. Chapter 2 discusses the utility of this "direction-to-direction" mapping for achieving motor-equivalent speech production, and presents the learning law. See Section 2.3.1 for the learning and performance phase equations. Those equations are adjusted to use the wavelet auditory representation, and are not repeated here.

The inverse kinematic map varies with articulatory configuration in a nonlinear manner. This variation is represented in the model by an adaptive neural network, which is described in the following section.

4.2.7 HRBF Approximation of Inverse Kinematic Map

A hyperplane radial basis function (HRBF) network is used to approximate the inverse kinematic map, which varies with articulatory configuration, for the simulations of DIVA reported in this chapter. HRBF networks were introduced by Stokbro, Umberger, and Hertz (1990).

Cameron (1996) provides a clear discussion of HRBF networks, which differ from ordinary RBF networks in two respects. The first difference between RBFs and HRBFs is that normalized Gaussian basis functions, $h_i = \frac{g_i}{\sum_j g_j}$, are used. The

second, and more important, difference is that the coefficient ν_i multiplying the basis function g_i in the expression $O = \sum_i \nu_i g_i$ is replaced by a "linear fit" term,

$$\nu_i + \sum_j c_i w_{ij}, \tag{4.1}$$

where $c_i = x - \mu_i$, and where μ_i is the center of the basis function. This linear fit term allows a hyperplane to be fit to the function, providing significantly greater accuracy than the standard RBF network. The number of free parameters for each basis function is increased by this method, but the number of free parameters scales linearly with the number of input dimensions, whereas the number of basis functions scales exponentially with the number of input dimensions. HRBFs have significantly better fit than RBFs for the same number of basis functions. For high-dimensional problems, the number of basis functions needed (and the corresponding memory and computation cost) is significantly smaller for a given total error. For example, it was determined that 162 HRBF basis functions are adequate to represent the inverse kinematic map with sufficient accuracy to produce the simulation results reported in this chapter.

Cameron (1996) developed Adaptive HRBFs and initially applied them to the problem of inverse kinematics approximations for speech production. A version of Cameron's computer software has been used in the simulations reported in this chapter and in Guenther et al. (1998), but with several differences. Although Cameron's gradient descent algorithm and software were utilized for this dissertation, his adaptive algorithm for selecting centers and widths of the basis functions was not, partly for simplicity, and partly because sufficient accuracy was achieved using fixed basis functions.

The number of trained parameters, or "weights", in the present model is 16 (8

basis functions, increase and decrease) times 14 (7 Maeda articulatory directions, increase and decrease) times 8 (free parameters per HRBF) times 162 HRBFs, or about 290,304 trained parameters for the inverse kinematic map. This is many more parameters than were needed for the formant planning model (see Section 2.3.1). However, the earlier model did not use HRBFs (which increases the number of free parameters by a factor of 8) and had only 2 planning dimensions (F1 and F2) compared to the 8 planning dimensions used here. In addition, the formant-based model was able to use only 54 tessellated regions, compared to the 162 HRBFs used here.

4.2.8 Phoneme-to-Auditory Map

The phoneme-to-auditory map computes the target region in the wavelet auditory planning space, given a desired phoneme. This map has the same form and obeys the same learning law as that of the phoneme-to-acoustic map defined in Chapter 2 for the formant-based version of DIVA, with the exception that the wavelet auditory representation (with 8 basis functions) is used instead of F1 and F2. Section 2.3.2 describes the learning equations, which have been adapted for use with the wavelet auditory representation.

4.2.9 Forward Model

The forward model allows computation of the wavelet auditory representation corresponding to a particular articulatory configuration, given by its Maeda parameters. Two different approaches to computing the forward model were explored. The first computes the wavelet auditory representation directly from the Maeda spectrum (the log magnitude Fourier spectrum computed from the area function resulting from the Maeda articulatory parameters. This approach is computationally intensive, but is the most accurate. The second approach uses the first to learn an HRBF network

representation of the forward model (i.e., a forward *map*). The HRBF network is described in Section 4.2.7. Although both approaches yield acceptable results, only simulations using the first approach are reported in this chapter. The method of computing the wavelet auditory representation is discussed in Section 3.4.2.

4.2.10 Babbling Phase

The present model employs a self-organizing neural network which allows the model to adapt to changing articulatory characteristics, and to learn the highly nonlinear relationships between the phonemic targets, their corresponding auditory/acoustic properties, and the necessary articulatory movements needed to realize them, without the need for explicit programming. A babbling phase generally consists of articulatory movements (random or otherwise) designed to allow efficient learning of the neural parameters. The model employs two distinct stages of babbling. In the first stage, the inverse kinematic map is learned. In the second stage, the target map is learned. It is possible that both maps can be learned simultaneously, although no attempt was made to accomplish this. There is evidence that humans employ different stages of babbling (see Appendix C for a short review of literature on human vocal babbling), although the connection between these stages and those used in the model has not be explored.

The neural network parameters, i.e., the phoneme-to-auditory mapping and the inverse kinematic (or auditory-to-articulatory) mapping, are learned during the initial babbling phase. For the simulations reported in this chapter, the following babbling algorithm is used.

Phoneme-to-auditory map: For each babbled movement, a random articulatory position is chosen from the space of possible Maeda articulatory positions, i.e., from the hypercube $-3 < x_i < 3$, where x_i is the articulatory parameter corresponding to

Table 4.2: Number of each vowel produced during a typical babbling session consisting of 50,000 babbled utterances.

/IY/	/I/	/E/	/AE/	/A/	/OW/	/U/	/OO/	/UH/
6	10	8	3	35	6	76	34	58

Table 4.3: Number of each stop consonant produced during a typical babbling session consisting of 50,000 babbled utterances.

/b/	/d/	/g/
69	29	104

the ith Maeda articulator. Illegal articulatory configurations, i.e., those corresponding to a vocal tract shape in which the vocal tract is completely closed off, are not used during learning.

During one typical phoneme babbling session, 50,000 babbled utterances were employed, 28,524 of these resulted in "legal" vocal tract shapes, and only 438 corresponded to recognized phonemes. Babbled utterances are obtained by selecting articulatory configurations randomly (using a uniform distribution over the space of articulator positions). Although the model selects movements randomly for babbling, it is unlikely that human infants use purely random movements during canonical babbling (see Appendix C). Tables 4.2 and 4.3 report, respectively, the number of each vowel and consonant produced during the babbling session. It can be seen that some phonemes are produced more often than others during babbling, reflecting the relative difficulty of reaching some articulatory configurations using the Maeda articulatory system, as well as the relative sizes of each of the phonemes in articulatory space. As long as each phoneme is produced a sufficient number of times, the corresponding auditory target is learned adequately.

Inverse kinematic mapping: The learning of the inverse mapping requires that *changes* in the articulatory position be paired with resulting *changes* in auditory

state. In order to obtain the articulatory changes or movements, a random articulatory position is chosen in the same manner as for the phoneme-to-auditory babbling described above. After this initial position is determined, an articulatory position nearby (at a distance of 0.6 units in articulatory space) is chosen. The direction of this movement is chosen randomly from among all possible directions. If the movement results in an illegal vocal tract configuration, then another direction is chosen at random, until a legal configuration is obtained. During numerous babbling sessions, there were no observed instances when a legal vocal tract configuration was not obtained by this method. Learning of the inverse map using this movement is described in Section 2.3.1.

4.3 Computer Simulation Results

This section summarizes results of computer simulations of DIVA modified to use the wavelet auditory representation of static vowel and stop consonant spectra. Some of these results were originally reported in Johnson (1997). Computer simulations have been carried out on a Sun SPARC-10 workstation running the SunOS 4.1.3 operating system. The computer simulation software is written in the C programming language. Graphical representations of articulatory configurations are generated by Xlib and X/Motif. Plots of Fourier and wavelet-smoothed spectra, and of trajectories in the F1-F2 plane, are produced by gnuplot. Programs are compiled and optimized using the gcc compiler, and were debugged with gdb.

4.3.1 Convergence of Initial Spectrum to the Target Spectrum

One consequence of planning articulatory movements in the wavelet auditory space instead of formant space is that the identity of formant peaks is not necessarily pre-

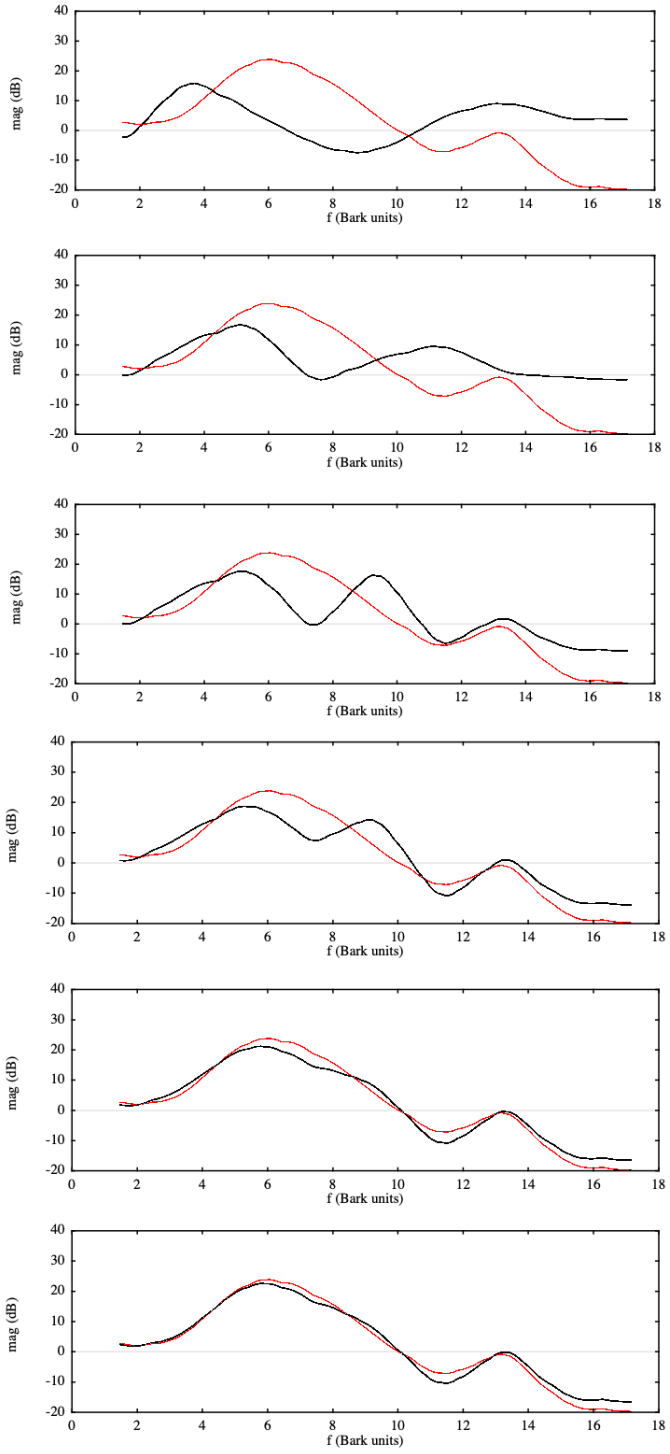


Figure 4·7: Convergence of spectrum (shown in black line) toward the spectral target for the /OW/ vowel sound (shown in red).

served during the production of vowel sounds. Consider the production of the /OW/ vowel sound. Figure 4·7 illustrates the sequence of intermediate spectra obtained during the production of /OW/, where the evolving spectrum is shown in the black line and the /OW/ spectral target is shown in red. Desired changes in the spectrum are mapped by the DIVA model to planned movements of the Maeda articulators, and these movements are reflected in the evolving spectrum.

In the top plot of Figure 4·7, both the starting spectrum, corresponding to the neutral configuration, and the /OW/ target have two peaks. By the third plot in the sequence, three peaks are evident in the evolving spectrum. In the fifth plot, the second peak has nearly disappeared. By the end of the utterance (plot 6 of Figure 4·7), the wavelet auditory spectrum has converged to the two-formant target in wavelet auditory space for the /OW/ sound.

This ability of the model to adjust the number of spectral peaks is a necessary feature of a wavelet-smoothed spectral representation of vowel sounds during speech production, where smoothing of the spectrum sometimes combines nearby spectral peaks (i.e., the spectral center of gravity effect). This feature is a consequence of the fact that the model uses just enough spectral resolution to recognize and produce vowel sounds, but not enough spectral resolution to represent separate formant peaks in every instance, when those peaks are more closely spaced than the resolution can resolve. Therefore, one of the claims of the model is that the wavelet auditory representation provides an optimal amount of spectral resolution for speech perception and production.

4.3.2 Vowel Production Results

Simulations of vowel production using the wavelet auditory planning space were designed to test the model's ability to produce each of 9 English vowels from a variety

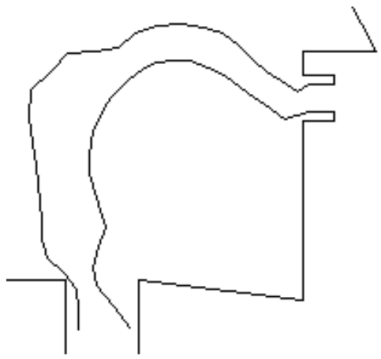


Figure 4·8: Neutral configuration of the Maeda articulatory system.

of starting vocal tract configurations. Initially, the system is reset to the neutral configuration (in which each Maeda articulator is set to zero), and then the system is commanded to produce the desired vowel. At the conclusion of each vowel utterance, the system is reset to the neutral configuration in preparation for the next vowel. Figure 4·8 illustrates the neutral configuration of the Maeda vocal articulators.

The trajectories in formant space produced by the model for each of the vowel utterances is shown in Figure 4·9. It should be emphasized that the wavelet auditory planning space is used to plan the speech movements, and that the formant plane is used only to illustrate the results of the simulation. The starting point of each trajectory is located at the formant pair (F1=431 Hz and F2=1778 Hz) corresponding to the neutral configuration. In each case, the end point of the trajectory is inside the formant criterion rectangle, at which point the vowel utterance ends. Although the trajectories produced by the model are not straight lines in the F1-F2 plane, and are probably not straight lines in wavelet space, the distance to the target rectangle decreases nearly monotonically, indicating that the inverse kinematic map is sufficiently accurate in those regions of articulatory space.

Two of the trajectories shown in Figure 4·9 deserve additional comment. The first, /IY/, is not as smooth as other vowels, especially near the end of its trajectory,

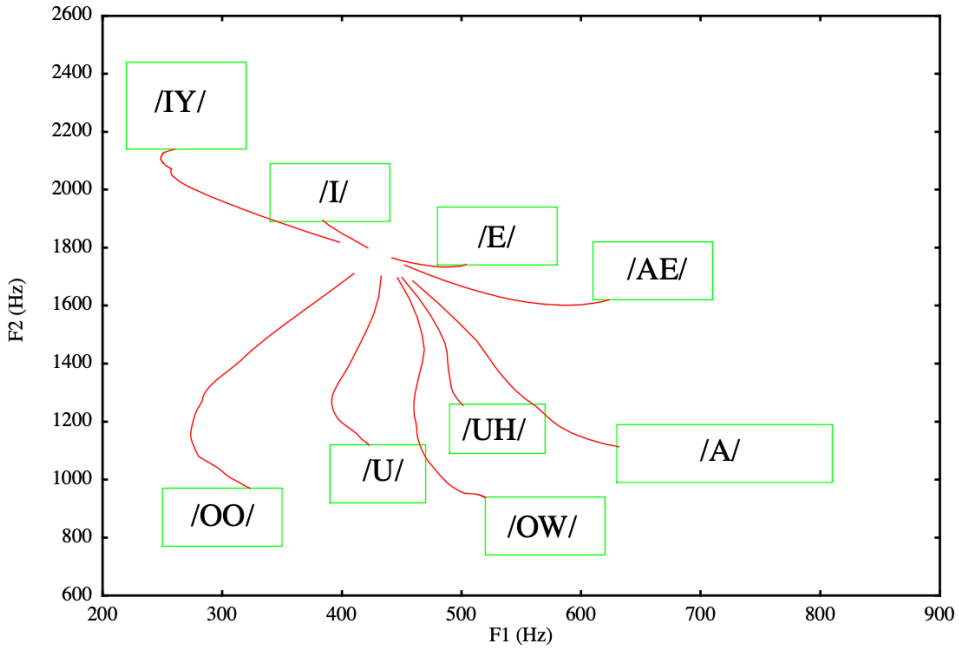


Figure 4·9: Trajectories produced by the model for all 9 vowels. Trajectories in wavelet planning space have been projected onto the F1-F2 plane.

where it approaches the target very slowly. This is probably because the wavelet-smoothed spectrum is already very close to the target spectrum, especially below about 2000 Hz, even though the value of F2 is not within the formant rectangle used for recognition of /IY/. This is an example of the formant uncertainty discussed in Section 4.2.3. However, the small difference between the current spectrum and the wavelet target continues to drive articulatory movement until the system eventually reaches the target, indicating that the inverse kinematic map has been adequately learned in that region of articulatory space.

The second, /E/ (and to a lesser extent, /AE/), could have been produced earlier in the trajectory had the value of F2 been increased earlier. Instead, F1 increases more than is apparently necessary, and then F2 increases near the end of the trajectory in order to bring the system inside the formant rectangle. However, it must be

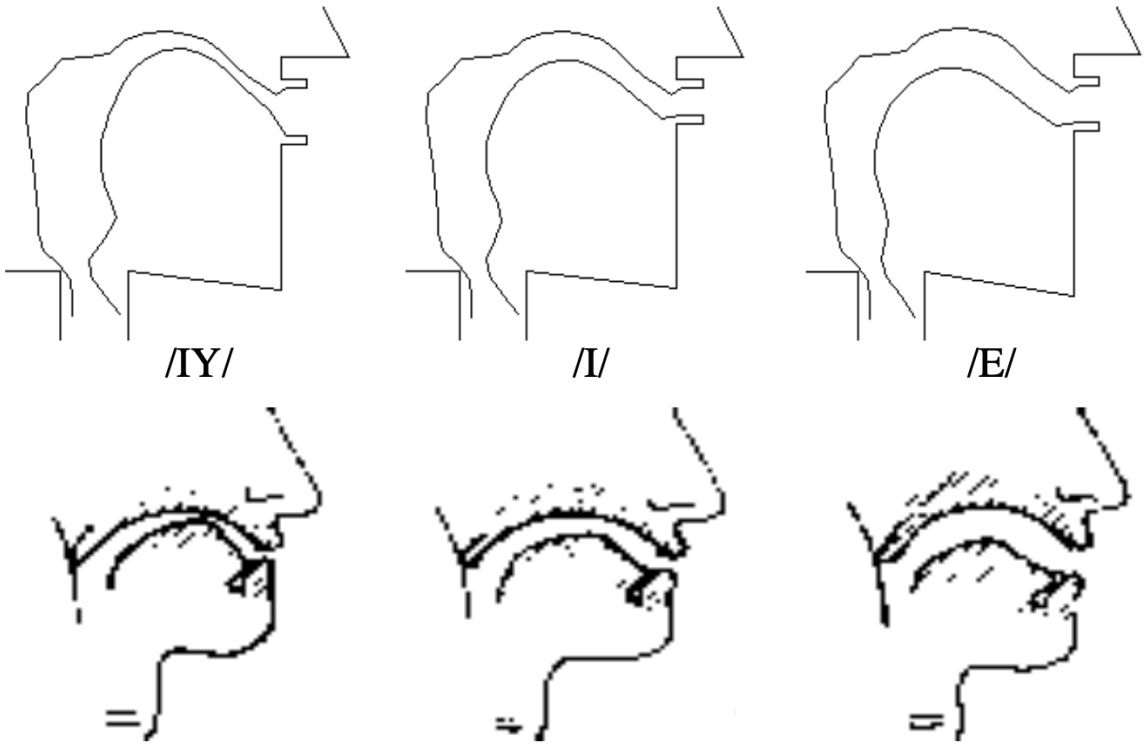


Figure 4·10: Typical vocal tract configurations produced by the model for /IY/, /I/, and /E/.

remembered that the articulatory movements are being planned in wavelet space, not formant space. At each time step, the model computes the movement that is most likely to decrease the distance to the target region in *wavelet* space. Moreover, the "target" rectangle in the F1-F2 plane is not identical to the hyper-rectangular wavelet target region. The formant "target" is not used at all by the model to compute desired movements. Instead, it is only used by the phoneme recognizer to decide when to terminate the utterance.

The model is successfully able to produce each of the 9 English vowels. The final vocal tract configuration corresponding to each vowel is shown in Figures 4·10 - 4·12. In these figures, the resulting Maeda articulatory configuration is shown along side the corresponding human vocal tract shape for comparison.

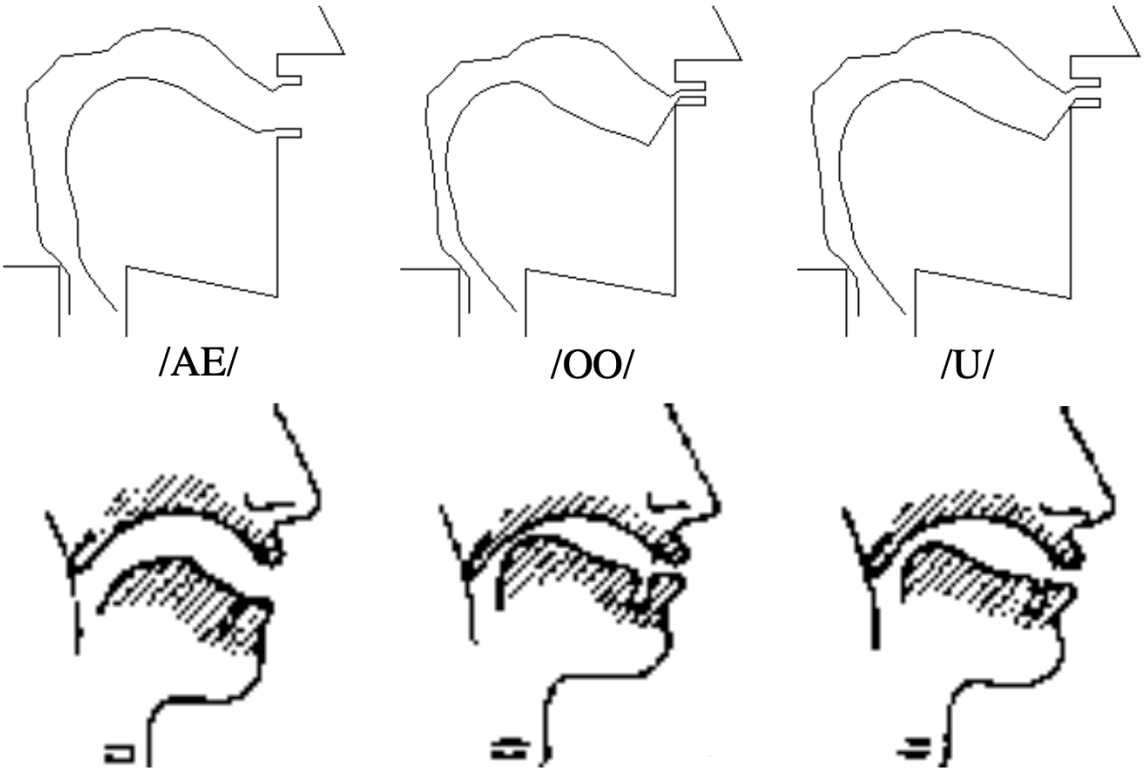


Figure 4·11: Typical vocal tract configurations produced by the model for /AE/, /OO/, and /U/.

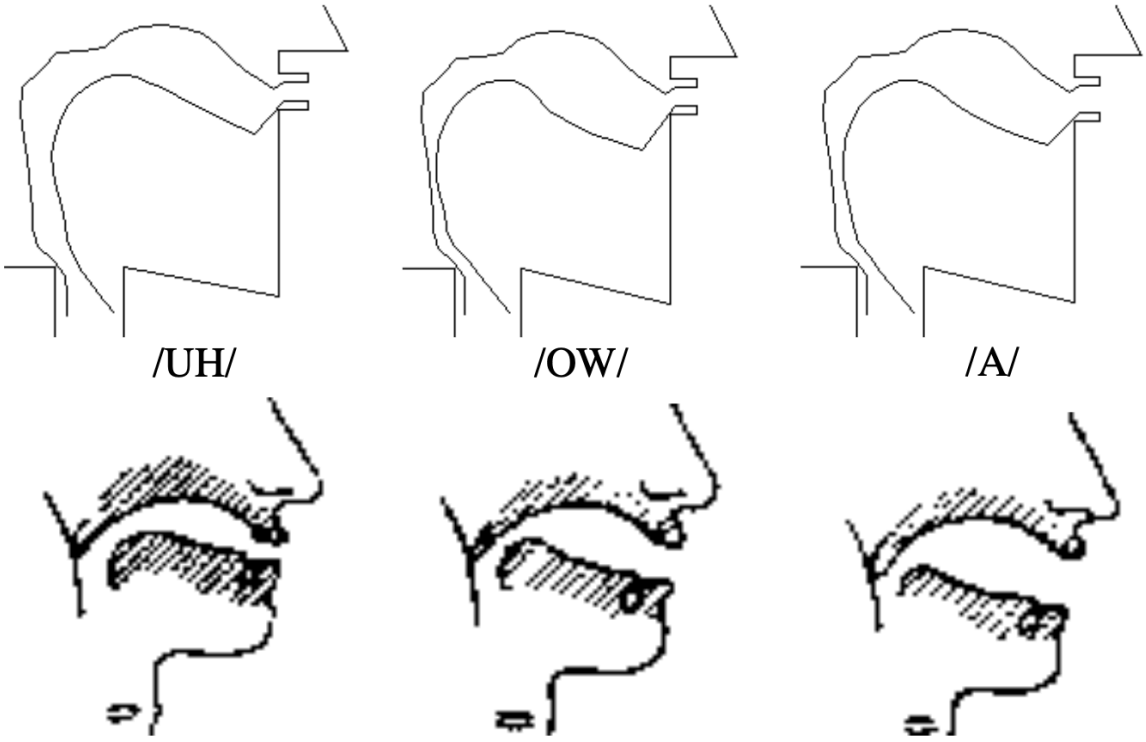


Figure 4·12: Typical vocal tract configurations produced by the model for /UH/, /OW/, and /A/.

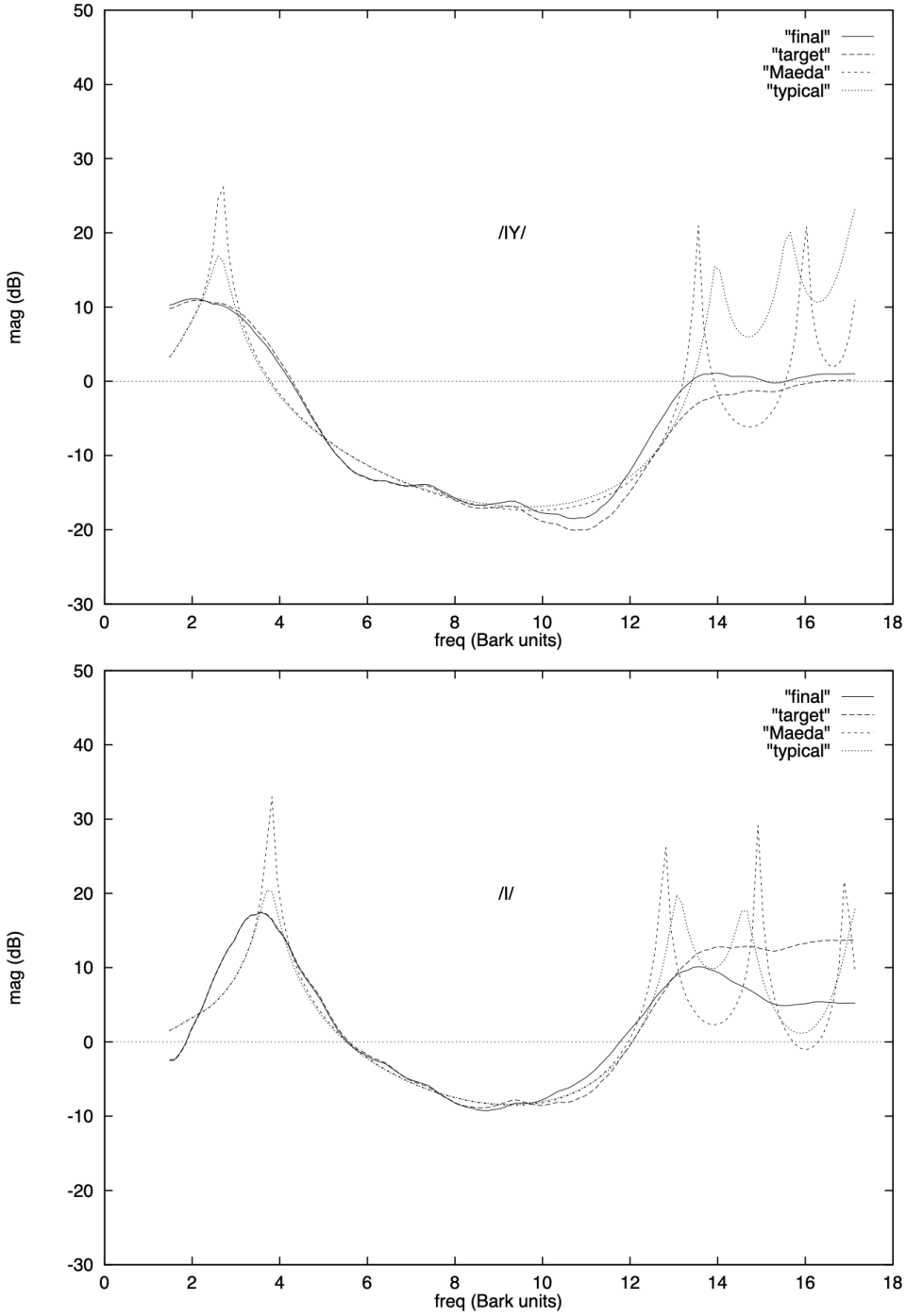


Figure 4·13: Ideal and smoothed vowel spectra shown for /IY/ and /I/.

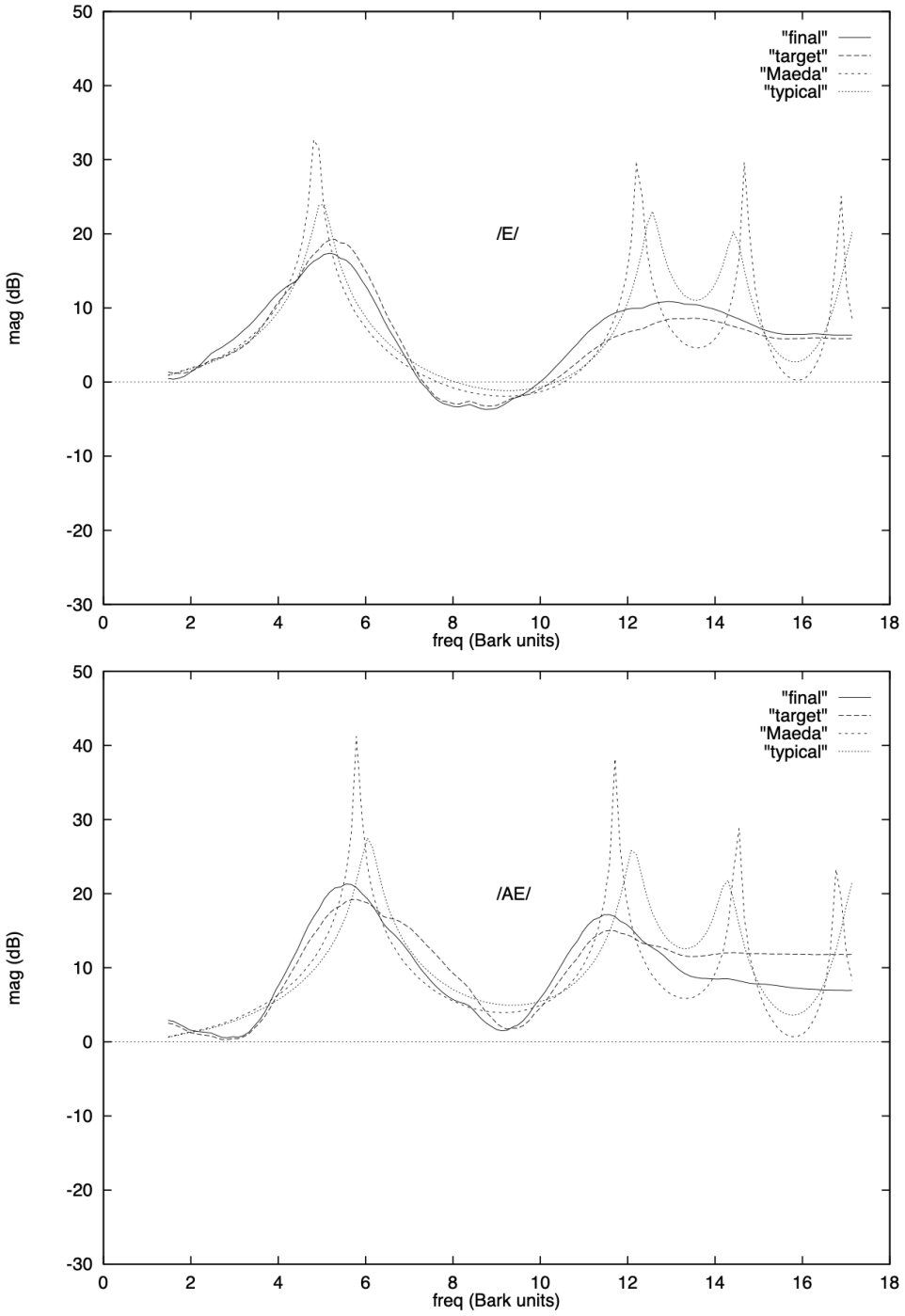


Figure 4·14: Ideal and smoothed vowel spectra shown for /E/ and /AE/.

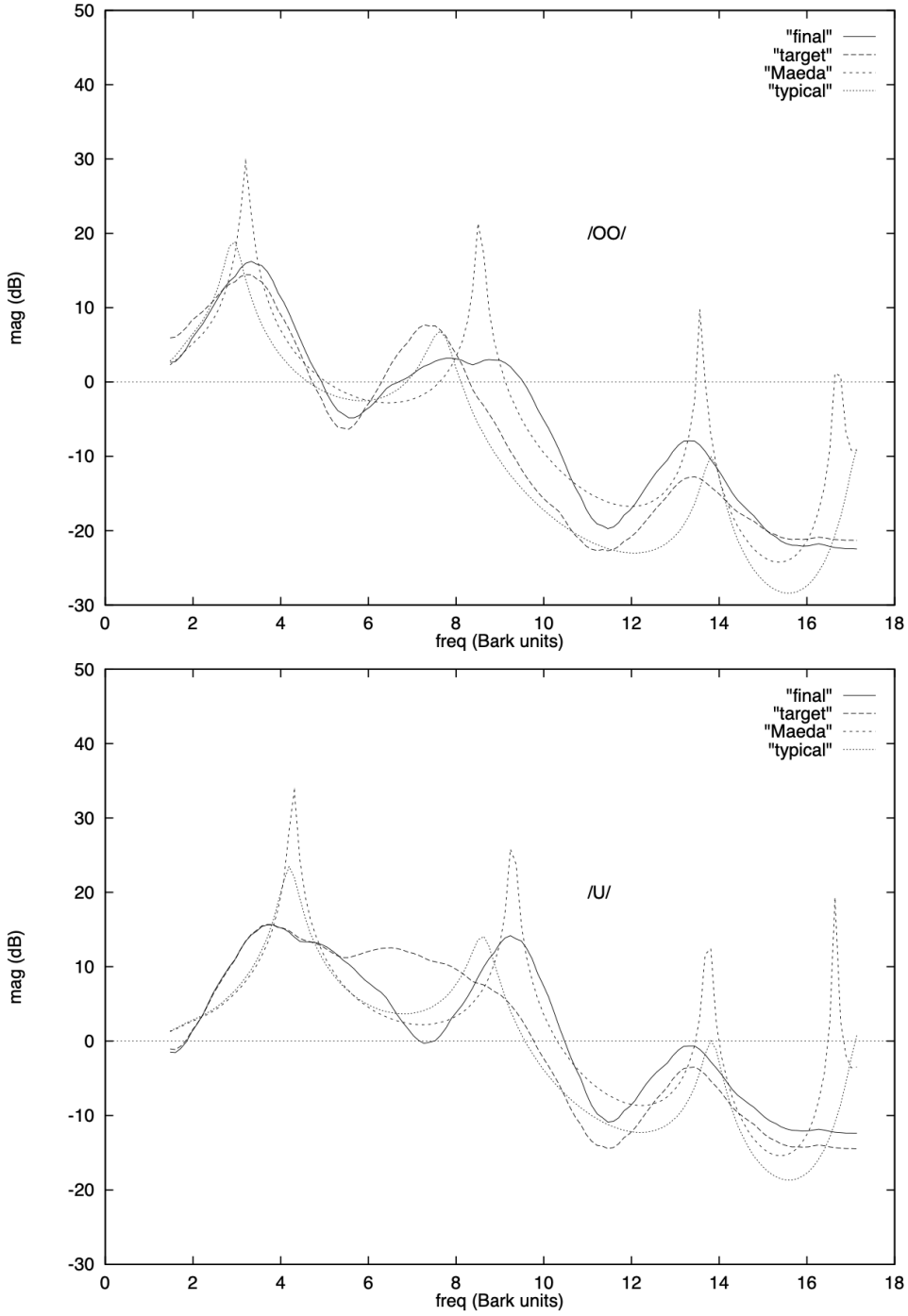


Figure 4·15: Ideal and smoothed vowel spectra shown for /OO/ and /U/.

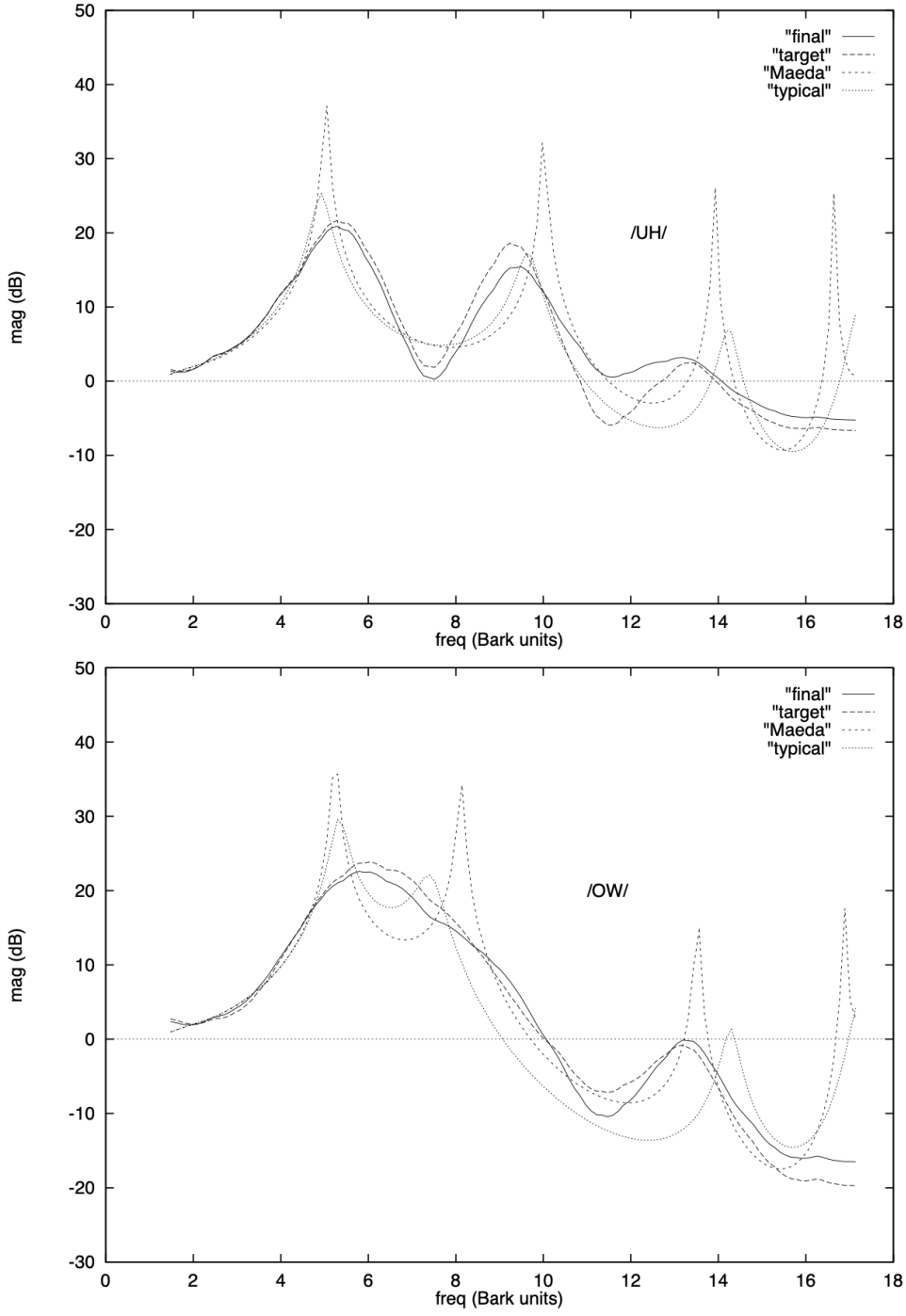


Figure 4·16: Ideal and smoothed vowel spectra shown for /UH/ and /OW/.

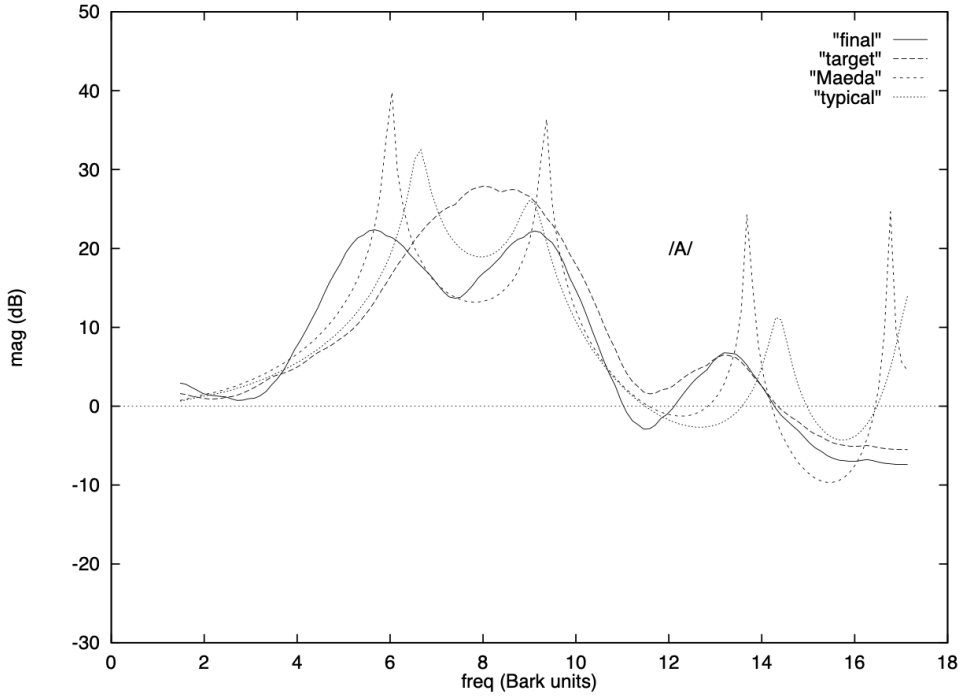


Figure 4·17: Ideal and smoothed vowel spectra shown for /A/.

The corresponding spectral results for each of the 9 vowels are given in Figures 4·13 - 4·17. In each plot, 4 curves are shown. The first is the "final" wavelet auditory spectrum produced by the model. The second is the "target" wavelet auditory spectrum for the given vowel. The third is the "Maeda" spectrum corresponding to the final articulatory configuration of the utterance. The fourth is the "typical" spectrum produced by the spectral generator described in Appendix A, in which F1, F2, and F3 frequencies and bandwidths, taken from Rabiner and Schafer (1978), are used to produce a typical spectrum for the given vowel for comparison with the Maeda-computed spectrum. In most cases, the final spectrum is very close to the target spectrum, usually differing significantly only at the higher frequencies. However, in the cases of /OO/, /U/, and /A/, there are larger differences between the final spectrum and the target. But in each of these cases, the final position in formant space

is within the criterion formant rectangle, which possibly indicates that the formant criterion rectangle is significantly larger than the target hyper-rectangular region in wavelet space. Another possibility is that the vowel is being produced using a vocal tract configuration which is farther from the neutral configuration than that chosen for deriving the wavelet "typical" target. In most cases, the Maeda spectrum is very similar to the typical spectrum. Differences in the amplitudes of the peaks between the Maeda and typical spectra result mostly from the fact that typical values of the bandwidths have been used in deriving the latter spectrum. Minor differences in the formant frequencies result from the fact that the typical spectrum uses values of formant frequencies at the *center* of the formant rectangle, whereas the Maeda spectrum is calculated from an articulatory configuration which occurs at the *boundary* of the formant rectangle.

4.3.3 An Additional Advantage of Using Region Targets

Region spectral targets may provide an advantage over point spectral targets by allowing more degrees of freedom of movement, especially when the system is near the vowel target. During preliminary investigations with the wavelet auditory planning space, point vowel targets were employed. There were several vowels that were difficult or impossible to produce, especially when the system was near the target. In particular, the lower half of the spectrum (below about 2 kHz) reached the target spectrum (corresponding to a single point in spectral space), but was unable to reach the target at higher frequencies. It is believed that the wavelet basis functions that code lower frequencies of the spectrum (the lower 4 basis functions from the set of 8) learned larger weights projecting to articulatory movements than the corresponding weights for the high-frequency basis functions. Therefore, a very good fit between the current spectrum and the target spectrum was obtained below 2 kHz, but unaccept-

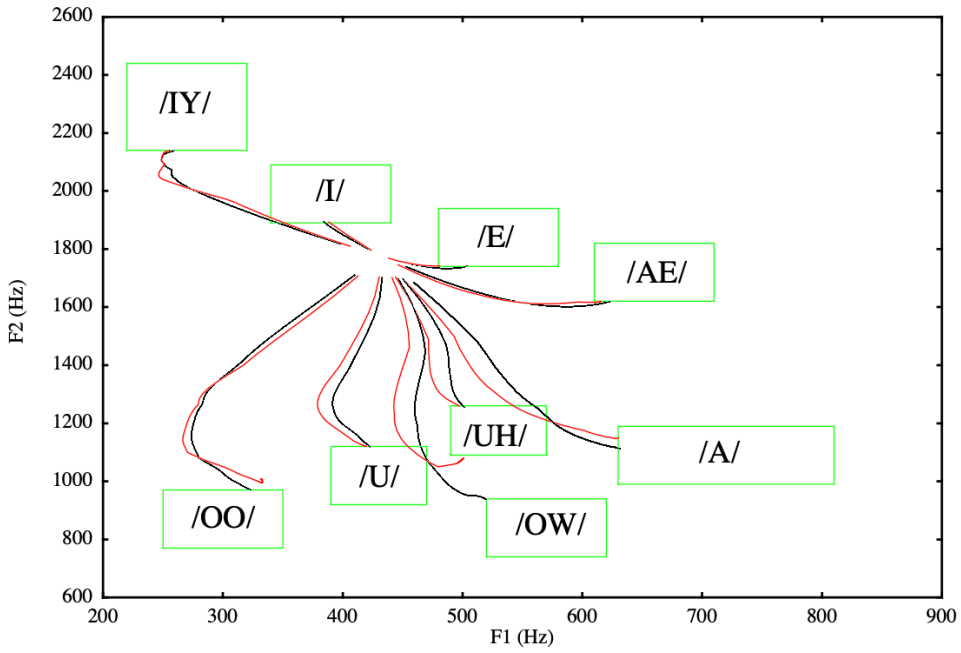


Figure 4·18: Trajectories with jaw blocked (shown in red) and with jaw unblocked (shown in black), for all 9 vowels. Trajectories in the wavelet planning space have been projected onto the F1-F2 plane.

able fits were obtained above about 2 kHz.

The problem was fixed by utilizing target *regions*. With a target region in the DIVA model, when the system is inside the target interval along one of the dimensions in the planning space, movements are inhibited for that planning dimension. This frees up the articulators to assist the system to reach the target intervals along the other dimensions of the planning space. Simulations using the wavelet auditory planning space provide anecdotal evidence for this idea. Vowels which were difficult or impossible to produce with point targets were easily produced from most starting configurations when region targets were used, even small regions.

4.3.4 Motor-Equivalence Results

In Chapter 2 (see Section 2.4.2), the motor-equivalence properties of the DIVA model using formant planning are illustrated. Similar results are obtained for the wavelet planning space, and these results are illustrated by considering the trajectories taken by the model with the jaw blocked, and by comparing those trajectories to the unblocked case. These trajectories are plotted in Figure 4·18, where black line trajectories correspond to the unblocked case and red line trajectories correspond the case of the blocked jaw. Although the trajectories are somewhat different for each of the vowels, for all but two of the vowels (/OO/ and /OW/), the targets are reached and the end-points are nearly identical, thereby demonstrating that motor-equivalent vowel production can be obtained when articulatory movements are planned in the wavelet auditory space.

After the failure to produce /OW/ with the jaw blocked at zero, the jaw was manually unblocked (allowed to move freely), and the model quickly lowered the jaw and successfully terminated the utterance inside the formant rectangle. Therefore, jaw lowering seems to be required for production of /OW/ by the Maeda articulatory system. After the failure to produce /OO/ with the jaw blocked, unblocking the jaw did not result in successful production of /OO/, suggesting either that the spectrum was already so close to the wavelet target that no further movement was planned, or that the inverse kinematic map was not adequately learned in that region of articulatory space. However, the /OO/ target was nearly reached with the jaw blocked. These results indicate that the DIVA model using the wavelet auditory planning space exhibits motor-equivalence during vowel production.

4.3.5 Consonant Production Results

Stop consonants are characterized by rapid movements of the speech articulators that result in rapid changes in the acoustic spectrum. Formant transitions have been used as a basis for stop consonant perception of place of articulation, but the simplest version of such a model does not work well because these formant transitions are sensitive to surrounding vowel context (Öhman, 1966). Sussman, McCaffrey, and Matthews (1991) discuss other recently proposed metrics, including release burst, VOT (voice onset time), shape of onset spectra, and locus equations. Stevens and Blumstein (1978) defined the onset spectrum of a voiceless stop consonant to be the short-time Fourier transform of the speech signal using a window which includes the release burst, the interval of silence, and the first glottal pulse of the following vowel. They derived distinct spectral shapes corresponding to the three stop place categories and argued that these onset spectra should be relatively independent of vowel context. Then they measured onset spectra for speakers of English, and found agreement with the derived spectral shapes (Blumstein & Stevens, 1980). However, Sussman et al. (1991) cite later research which casts doubt on the role for onset spectra in stop consonant perception, especially in languages other than English.

Instead, they suggest that the so-called locus equation metric is the most predictive of stop place of articulation independent of following vowel context. A locus equation is a linear regression of F2 onset frequency on the F2 vowel frequency, and they have shown that the slope and y-intercept predict place of articulation robustly across gender. Furthermore, the standard error of the regression is very low, i.e., the linear regression is a very good fit of the data. For a summary and review of the locus equation idea, see Sussman, Fruchter, Hilbert, and Sirosh (1998).

Combination-sensitive cells, capable of detecting linear relationships within an

acoustic signal that are similar in some respects to locus equation relationships, have been found in the auditory systems of the mustached bat, brown bat, cat, mouse, monkey, ferret, and frog. Mendelson and Cynader (1985), for example, report that some cells in cat auditory cortex are tuned to one particular direction of frequency modulation and some are tuned to a certain range of frequency modulation. Others have shown that there exist cells that are tuned to amplitude-modulated stimuli (Schreiner & Urbas, 1986). Shamma et al. (1995) studied the temporal properties of cells in ferret primary auditory cortex, and found that temporal features are coded, in part, by basis functions with an asymmetrical spatial receptive field shape. In particular, they considered the response of these asymmetrical cells to swept FM tones. Up and down sweeps were used, and their responses were measured. It was found that cells with greater inhibition at higher frequencies responded better to upward sweeps, and cells with greater inhibition at lower frequencies responded better to downward sweeps. Sussman et al. (1998) cite other studies which lend support to the idea that such combination-sensitive cells are pervasive among animals and may play a role in explaining the locus equation in stop consonant perception and production.

These considerations suggest that static spectral targets are insufficient for the production of stop consonants that can be easily recognized by human listeners. The present model does not provide the necessary control over articulator velocity, or, equivalently, the velocity of the trajectory in the planning space. Therefore, no attempt has been made in the present research to replicate locus equation results or the complex temporal dynamics seen during human consonant production. However, as was mentioned in Section 4.2.5, the static spectral targets corresponding to the voiced stops have been determined and have been used to control the vocal articulators to

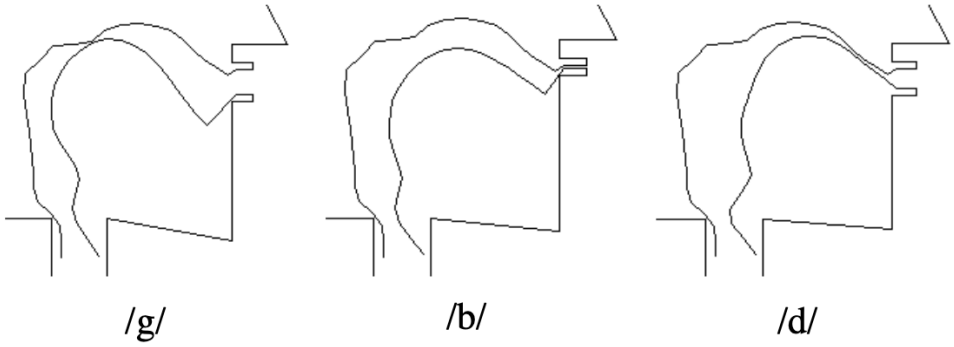


Figure 4·19: Vocal tract configurations showing constrictions corresponding to voiced stop consonants /g/, /b/, and /d/.

produce the static vocal tract constrictions corresponding to the three voiced stop consonants, /d/, /g/, and /b/.

In the case of two of the stops (/g/ and /b/), the correct vocal tract constrictions are obtained by the model without any extra manipulation. These configurations are shown in Figure 4·19. For /d/, correct production required preventing lip aperture movement during the utterance. If lip aperture movement is prevented, the tongue body moves until a constriction in the dento-velar region, consistent with production of /d/ in humans. However, the fact that lip aperture must be explicitly restricted in order to produce /d/ suggests that spectral information in the form of the wavelet auditory representation, in the absence of constriction degree and location information or additional acoustic information, may not be sufficient to distinguish between /b/ and /d/ during production of these stop consonants.

4.3.6 Why a Log Frequency Scale in Auditory Cortex?

Results of experiments with linear and Bark frequency scales in the present model suggest that a better inverse kinematic map is learned using the Bark scale. The choice of scheme for coding frequency, (e.g., linear, logarithmic, or Bark) affects the stability and training time of the system in the following manner. Öhman (1966) quotes Heinz

and Stevens (1964) in claiming, "The dependence of a formant frequency on small irregularities in the shape of the vocal tract increases rapidly with the ordinal number of the formant" (p. 163). The formant frequency increases with its ordinal number. Thus, the spectrum has higher variability, for a given change in the vocal tract area function, at the higher frequencies. Small changes in articulators cause small changes in the spectrum at low frequencies but large changes at high frequencies.

With a linear frequency scale, the wavelet representation employs the same number of basis functions per unit frequency at low and high frequencies. Thus, the frequency resolution of the representation is the same at the two extremes. This results in more frequency resolution in the spectral representation than is necessary at low frequencies, but less frequency resolution than required at the higher frequencies, making learning of the inverse kinematic map more difficult at both ends of the frequency spectrum. It is more difficult to learn at the high end because the representation does not have enough resolution there to represent the spectrum with sufficient accuracy. It is more difficult to learn at the low end because there is an excess number of basis functions, implying that any one of them will exhibit more variability than can be reliably learned.

Studies of primary auditory cortex by Shamma suggest that the frequency axis is linear to about 500Hz, then logarithmic (base 2) beyond 500Hz. In modeling of primary auditory cortex, Wang and Shamma (1994) use a log frequency axis, but they do not assume linearity below 500Hz. The Bark scale uses a similar coding scheme. Zahorian and Jagharghi (1993) also show that vowel recognition (using automated classifiers) is optimal using log amplitude and Bark frequency scaling. Their results hold for both formants and DCTCs (discrete cosine transform coefficients).

A representation that equalizes the variance across each of the basis functions is

most desirable, since it maximizes utilization of these basis functions. Basis functions that code zero variance are wasted in the sense that they are unnecessary. There is no point in having a representational scheme to store a formant frequency or amplitude that never changes, for example. On the other hand, basis functions that code high variance are probably not going to be very well learned. In that sense, they are also wasted because they will not be able to do the job, having too much error. The problem of unequal distribution of variance is one of the problems with a formant representation. The lowest two formants are relatively stable against variations in vocal tract shape, but the higher formants are much more unstable. However, a spectral shape representation with the appropriate frequency scaling (e.g., logarithmic or Bark) will distribute this variance uniformly and will improve learning of the inverse kinematic mapping used in the model.

4.4 Discussion and Conclusions

This chapter began by describing preliminary investigations with the wavelet auditory representation of vowel spectra which suggested that linear interpolation between vowel targets in this space result in acceptable vowel transitions and final vowel sounds of acceptable sound quality. The remainder of the chapter discussed the modifications to DIVA necessary to use the wavelet auditory planning space in vowel and stop consonant production, and reported simulation results using this planning space. These results show that the proposed model explains many of the static properties of speech. In particular, the proposed model successfully produces each of 9 English vowels from a variety of starting vocal tract configurations. Evidence for this conclusion include plots of trajectories in the formant plane, graphical representation of final vocal tract shapes, and acoustic output that can be easily recognized as the

appropriate vowel sounds. In addition to vowels, the model is also able to produce the static vocal tract constrictions associated with the voiced stop consonants /d/, /g/, and /b/. However, the model has some difficulty distinguishing between /d/ and /b/, suggesting that additional acoustic information or some information about constriction degree and location may be necessary for the planning of stop consonant production. In addition to the ability to produce vowels, the proposed model is also able to produce most of these vowels with a blocked jaw, thereby demonstrating the motor-equivalent capabilities of the model.

These vowel and stop consonant production results, taken together with the results from the previous chapter on the spectral center of gravity effect and the physiology of primary auditory cortex, suggest that the wavelet auditory representation of vowel spectra is an appropriate representation for both speech production and perception.

Chapter 5

Conclusions

The research reported in this dissertation extends the earlier results with the DIVA model of speech production by showing that a physiologically plausible representation of the vowel spectrum, thought by many to exist in primary auditory cortex, can be used by the brain's motor system to control the movements of speech articulators. This chapter first summarizes the results presented in earlier chapters. Then several natural extensions of this research are suggested for future study.

5.1 Contributions of the Thesis

The main contributions of this dissertation include the following:

- Proof is given that acoustic planning of speech production using formant frequencies is possible. DIVA is the first computational model of speech production that is capable of producing correct vocal tract shapes during vowel production using acoustic planning, without using explicit constriction targets. This is a very significant result because of the recent evidence that suggests that humans plan vowel and consonant production in an acoustic-like space. The simulation results also show that the motor-equivalent features of the DIVA model are enhanced when an acoustic planning space is used.

- A new representation of vowel spectra, the wavelet auditory representation, is

proposed, inspired by the physiological modeling of primary auditory cortex by Shamma. Although the proposed model does not employ a full multiscale representation of the spectrum, Shamma (1988) suggests that only the largest scales are required to represent the gross spectral shape. The smaller spectral scales might be utilized to extract and represent the pitch of the utterance. The existence of this kind of representation in the human brain is also supported by substantial psychophysical results which show that gross spectral shape is more closely correlated to vowel perception. The wavelet auditory representation provides a means for representing and computing the gross shape of static speech spectra such as those seen in vowel production. Peaks in the wavelet spectrum sometimes, but do not necessarily, correspond to formant peaks. The wavelet auditory representation is simpler to compute than the formant representation.

- A new explanation of the spectral center of gravity effect, based on the wavelet auditory representation, is given. The spectral center of gravity effect is the phenomenon in which spectral peaks closer than about 3 Bark units are averaged into a single peak for the purpose of vowel perception. A formant representation of the vowel spectrum is unable to account for this phenomenon. By fitting the data from experiments on the spectral center of gravity effect, this thesis shows that it is possible to constrain the bandwidths of the wavelet basis functions in the wavelet auditory representation. The resulting representation employs 8 orthonormal basis functions that span the space of log magnitude short-time Fourier transform spectra between 0-4000 Hz.

- A new planning space for the production of vowels and consonants, based on the above-mentioned wavelet auditory representation, is presented. The DIVA model is modified to employ the wavelet auditory planning space, and numer-

ous simulations confirm that the modified DIVA model enjoys all of its earlier advantages over prior models, including production of vowels with correct vocal tract shapes without explicit constriction target information, compensation with blocked articulators, and the ability to self-organize during a babbling phase. The representation of the wavelet auditory planning space is distributed and physiologically plausible, as opposed to a formant planning space.

5.2 Areas for Future Work

A number of possible areas for future research include the following:

- The simulation results of the previous chapter utilized only two configurations of the wavelet transform. Each configuration employed 4 wavelet levels and 1 scaling level. One configuration employed 128 basis functions, resulting in 8 scaling functions in the wavelet auditory representation, and the other configuration employed 256 basis functions, resulting in 16 scaling functions in the wavelet auditory representation. There are many other possible configurations of the discrete wavelet transform that might give different simulation results. One possibly very interesting area of research involves consideration of different numbers of wavelet scales, wavelet bandwidths, etc., in order to reduce the number of basis functions or increase the accuracy of speech production. For example, a configuration that uses fewer scaling functions (perhaps two), but more wavelet levels in the final wavelet auditory representation, may be able to produce similar results with fewer basis functions. It will be surprising if the 8-basis wavelet auditory representation proposed in this dissertation turns out to be optimal for both vowel perception and production.

- Future work should quantitatively analyze the goodness of fit between the ar-

ticulatory configurations obtained for vowels by the model and the range of articulatory configurations used by humans. The use of target regions in the acoustic planning space implies that a range of articulatory configurations will be obtained by the model. One very important theoretical question is whether the range observed in human utterances can be explained by the target region concept. A more thorough statistical analysis of the model results might answer that question.

- This dissertation has not explored, in any depth, the problem of vowel perception. The wavelet auditory representation provides a means for encoding vowel spectra that may be more suitable for learning vowel categories in a self-organizing classifier such as ARTMAP (Carpenter, Grossberg, & Reynolds, 1991b). In addition, the wavelet auditory representation may provide a natural framework for understanding the vowel perception results of Syrdal and Gopal (1986), in which the same vowels tend to cluster within a 3 Bark interval in formant space.

  Related research questions concern the issues of speaker-normalization and speaker-independence. Syrdal and Gopal (1986) argue that a representation based on the bark scale offers significantly better speaker normalization than formant ratios. The wavelet auditory representation, or a variant, may provide a speaker-independent representation of vowel sounds.

- One of the nice features of a representation based on gross spectral shape, particularly the wavelet auditory representation, is that it is likely to be more robust in noisy environments. This is because the wavelet decomposition of the spectrum provides a separate representation for small- and large-scale spectral

features, and because spectral noise constitutes a small-scale spectral feature which is thrown away when deriving the wavelet auditory representation.

- Another reason for questioning the role of formants for speech production is the ubiquity of curved trajectories in formant space. If humans use straight-line trajectories to plan vowel production in a formant space, then why are the resulting formant trajectories curved? That this occurs during human speech production can readily be demonstrated during production of diphthongs (Holbrook & Fairbanks, 1962). Results of DIVA simulations with the wavelet auditory planning space also exhibit curved trajectories in formant space, even though the inverse kinematic map was learned very well in all regions of articulatory space. Future research might seek to discover whether straight line trajectories in wavelet space correspond to curved trajectories in formant space, and if not, why curved trajectories occur in formant space. Also, although the inverse kinematic map was learned well enough to allow the model to produce all of the English vowels, residual errors in this map might still account for the curved trajectories in formant (and wavelet) space.

- This dissertation has focused on issues relating to vowel production and perception, and has ignored many of the much more complicated problems of consonant production. Future research might improve the modeling of consonant production, and include the simulation of the locus equations and some temporal aspects of consonant production. This modeling work might incorporate asymmetrical wavelet basis functions corresponding to cells thought by Shamma to exist in primary auditory cortex which demonstrate sensitivity to FM rate and direction. Such modeling work would generalize, in a natural way, the static spectral results provided herein. It would also address shortcomings with the

DIVA model with respect to control of the velocity of articulators.

- Although the hyperplane RBF (HRBF) networks employed in this dissertation provide very good results, it is believed that a carefully designed neural network based on adaptive wavelets (perhaps in conjunction with the idea of a hyperplane approximation) will produce better results and will be easier to learn. Preliminary results of work by this author (not reported in this dissertation) show that wavelet neural network approximations of the forward and inverse kinematic maps are feasible. Much more work is needed in this area.

The research presented in this dissertation only scratches the surface of the many interesting problems that remain in developing an understanding of speech production by humans.

Appendix A

A Simple Spectrum Generator

This appendix describes a simple algorithm for the generation of LPC spectra suitable for use in studying the spectral center of gravity effect.

Because the model does not treat nasalized vowels, it may be assumed that the vocal tract transfer function can be represented by poles in the complex plane. Thus, it has the form of the quotient of a constant and a polynomial which can be factored. Flanagan (1957) shows that a cascade speech synthesizer is adequate for nonnasalized vowels. This leads to the product form for the z-transform representation of the vocal tract transfer function. Following Klatt (1980), assume that each spectral peak can be represented by a second degree factor in the z-transform domain. Each factor can be thought of as a digital resonator. The resulting vocal tract transfer function can be written in the form:

$$T(f) \cong T(z) = \prod_{i=1}^{N} \frac{A_i}{1 - B_i z^{-1} - C_i z^{-2}} \tag{A.1}$$

where

$$C_i = -e^{-2\pi\delta_i T} \tag{A.2}$$

$$B_i = 2e^{-\pi\delta_i T} cos(2\pi F_i T) \tag{A.3}$$

$$A_i = 1 - C_i - B_i \tag{A.4}$$

and

$$T \equiv \text{period } (= 1/\text{sampling rate}) \tag{A.5}$$

$$\delta_i \equiv \text{bandwidth of the } i\text{th formant} \tag{A.6}$$

$$F_i \equiv \text{frequency of the } i\text{th formant.} \tag{A.7}$$

The z variable can be written

$$z = e^{j\omega T} \tag{A.8}$$

where

$$\omega = 2\pi f. \tag{A.9}$$

In the simulations reported in this dissertation, the sampling rate is 8000 samples per second, $T = 125$ microseconds, and formant bandwidths range from about $40 - 240$ Hz.

A spectrum with N formants can be obtained by setting the N values of F_i and δ_i, and computing $T(f)$ for the necessary values of f. The bandwidth of the ith formant, δ_i, is adjusted until the desired formant amplitude is obtained.

Appendix B

Mathematical Details of the Wavelet Formalism

Mathematical details of the wavelet formalism are presented in this appendix.

Given the "mother wavelet" $\psi(x)$ defined in Section 3.4.1 and having easily satisfied properties (Mallat, 1989; Daubechies, 1992; Kaiser, 1994; Vetterli & Kovačević, 1995), its integer dilates and translates constitute a set of orthonormal basis functions $\psi_{j,k}(x) = 2^{-j/2}\psi(2^{-j}x - k)$ which spans $L^2(R)$.

What is desired is a suitable linear decomposition of an arbitrary function f into a series expansion of the form

$$f = \sum c_i \psi_i, f \in L^2(R) \tag{B.1}$$

The Fourier transform is an example of such a decomposition, but in that case the $\psi_i \notin L^2(R)$, and the ψ_i do not have adequate time localization. Because of this lack of time localization, the Fourier transform is not suitable for representing signals whose spectral properties vary in time (such as speech). The STFT imposes time localization on the basis functions by multiplying the sinusoids by a windowing function. However, it is impossible to construct orthogonal bases for the STFT that also have "good" time-frequency localization properties (see Daubechies, 1992, Section 4.2.2). This problem led Morlet and Grossman (1983) to introduce the wavelet transform which uses constant-Q filters, i.e., the basis functions can be thought of as being constructed from sinusoids multiplied by $g(x)$ whose window length is proportional

to the wavelength. Thus, the basis functions all have the same shape.

Following the development of Mallat (1989), assume there exists a *scaling function* $\phi(x)$ such that the set of integer translates of ϕ, $\{\phi(x-k) : k \in Z\}$, are orthogonal. Such a ϕ can be shown to exist under suitable conditions. Let V^0 be a space of functions spanned by the translates of ϕ, i.e., $V^0 \equiv \text{Span}(\phi(x-k))$, $k \in Z$. V^0 is the space of all function approximations at scale 0. Given a function f, its closest approximation $f^0 \in V^0$ can be found by computing

$$f \approx f^0 = \sum_k c_k \phi(x-k), \tag{B.2}$$

where $c_k = \langle \phi(x-k), f \rangle$.

In addition to orthogonality, we will require that ϕ be refinable, i.e.,

$$\phi(x) = \sum_k a_k \phi(2x-k), \tag{B.3}$$

in other words, that the scaling function can be written as a linear combination of compressed copies of itself. It is this latter property which allows determination of possible functions $\phi(x)$, since it is a fixed point of Equation B.3.

Equation B.3 provides a connection between successive approximation spaces V^j and allows definition of a nested sequences of such spaces:

$$V^0 \subset V^1 \subset V^2 \cdots \tag{B.4}$$

By virtue of this refinement property, it can be shown that $(\sqrt{2^{-j}}\phi_j(x - 2^{-j}k))_{k \in Z}$ is an orthonormal basis of V^j, where $\phi_j(x) = 2^j \phi(2^j x)$. It should be noted: While the translates of ϕ_j are orthogonal within a scale, in general they are not orthogonal across scales.

Consider two adjacent spaces in the nested sequence, V^j and V^{j+1} and define W^j

such that

$$V^{j+1} = V^j \oplus W^j, \tag{B.5}$$

i.e., where W^j is the orthogonal complement of V^j in V^{j+1}. The space W^j is the *detail* space and codes differences between V^j (lower resolution) and V^{j+1} (higher resolution). Together V^j and W^j span the space V^{j+1}.

Finding a projection of an arbitrary function onto W^j requires that we know a basis for W^j. Such a basis can be determined in a manner analogous to that for the V^j, i.e., there exists a function ψ such that $(\sqrt{2^{-j}}\psi_j(x-2^{-j}k))_{k\in Z}$ is an orthonormal basis of W^j, where $\psi_j(x) = 2^j\psi(2^j x)$. Therefore, $(\sqrt{2^{-j}}\psi_j(x-2^{-j}k))_{j,k\in Z^2}$ is an orthonormal basis of $L^2(R)$.

Appendix C

Literature Review of Infant Vocal Babbling

The study of human infant vocal behavior has grown substantially in recent years and much is now known about the acquisition of speech and language in the early years of life. However, several outstanding questions remain. Among these is: What is the role of babbling in the infant? Models of speech in which babbling plays a central role have been proposed. However, these models typically ignore basic facts about infant babbling.

Babbling is a stereotyped behavior that precedes or accompanies motor skill acquisition in animals. Little is known about the mechanisms or role of babbling.

Roman Jakobson (1941/68) postulated that (1) babbling and meaningful speech are distinct processes, (2) babbling has astonishing diversity, and (3) babbling has little or no regularity (i.e., it is random). In the past two decades, many researchers have studied babbling in human infants and have found that some of Jakobson's ideas need to be reconsidered in light of the data.

While infants exhibit a diverse vocal repertoire during the babbling and first words phases, they also exhibit considerable regularity. Thus attempts to model babbling as a random process are not given support by the large literature on babbling. Moreover, the vocal repertoire during babbling, while large, is but a small subset of that seen in adult language (as opposed to the common view that babbled sounds are a superset of adult productions). Infants acquire new productions (phonemes) only very slowly

as the demands of word acquisition are imposed, and often make due with sounds from their babbling repertoire to create quasi-words, and imitate (as best they can) words from their target language. By the same token, the babbling literature also makes it clear that babbling and meaningful speech are not distinct processes. This is especially clear when considering studies of infants in the late babbling stage during which they are acquiring their first words. There is both phonological and phonetic continuity during this period, making it very difficult even to define separate stages of babbling and word acquisition.

On the question of environmental influence on the development of language, the literature provides mixed answers. Mother's speech plays a role, but cross-cultural differences do not seem to affect the onset of babbling or its content. Most significantly, linguistic environment does not seem to affect the babbling repertoire. The literature demonstrates that the babbling repertoire of infants from such divergent linguistic environments as English, French, Thai, Chinese, and Dutch all are phonetically very similar. Differences emerge only slowly, usually beginning with vowels, and only later are differences in consonants learned.

In general, the onset of babbling in normal infants is robust, occurring sometime between 7-10 months of age (always by 10 months). This is even true in mentally retarded infants, and premature infants. On the other hand, auditory feedback (or "tactile speech", in the case of acochlear infants), is required for the onset and continuance of babbling. But it is not clear whether the infant must hear the speech of others, or whether it is the infant's own speech that is required for the onset of babbling (See Locke & Pearson, 1990, on tracheostomized infants).

While our knowledge of babbling and early language acquisition have grown tremendously in recent years, many questions remain. Some of these questions are:

What is the role of feedback (visual, auditory, or tactile) in the onset of babbling and the development of speech? How does linguistic environment influence the late babbling and first words stages? To what extent are babbling and language innate? Perhaps the most important question is: What role does babbling serve in the acquisition of speech and language?

C.0.1 Pre-Canonical Babbling

Prior to the onset of canonical babbling (which occurs at 7-10 months of age in normal infants), the infant is capable of producing a variety of vocal sounds.

Koopmans vanBeinum and van der Stelt (1986) studies this stage in the development of the infant. 69 infants were studied based on written survey of their parents. Comfort sounds, which can be reliably identified by adults, are used to study infant vocalizations. Milestones are identified for phonation, articulator position, and articulator movement. An infant was assigned to a stage if two attributes of that stage were present. Most infants could be assigned to a stage every two weeks during the study. Koopmans vanBeinum and van der Stelt (1986) concluded that: (1) Each infant goes through six different stages. (2) By stage 5, all motor elements of adult speech are present, at least in rudimentary form. (3) Infants of the same age can belong to totally different developmental stages.

C.0.2 Canonical Babbling

This section discusses canonical babbling, the onset of canonical babbling, and variability of sounds during babbling, both within an infant, across infants within a single language environment, and cross-linguistically.

The term *babbling* is often associated with production of repeated CV syllables such as “bababa” and “dadada”. Such a production, (CV, CVCV, or CVCVCV)

is referred to as a *babble*. *Reduplicative babbling* refers to production of a babble in which the same (or nearly the same) consonants and vowels occur. In *variegated babbling*, the consonant or vowel differs during a babble. *Canonical babbling* usually implies either reduplicative or variegated babbling.

The study of canonical babbling is important for what it might teach us about speech, language, and developmental processes. The study of canonical babbling may also find application in the diagnosis of various kinds of speech disorder. One study (Stoel-Gammon, 1989) focused on two infants, out of 34 infants studied from 9 to 24 months, whose speech skills were below criterion levels at 24 months of age. It was found that these two "late talkers" had unusual patterns of babbling from 9 to 21 months of age. While Stoel-Gammon (1989) are cautious in their conclusions, they suggest that "[atypical babbling] may be *one* of many possible contributing factors" to the delay of language acquisition [emphasis in original].

C.0.3 Acoustic Features of Babbled Sounds

Many researchers have studied the acoustic features of babbled sounds (Mitchell & Kent, 1990; Kent & Murray, 1982; Davis & MacNeilage, 1990; Roug, Landberg, & Lundberg, 1988; Elbers, 1982).

For example, Roug et al. (1988) studied place of articulation, manner of articulation, degree of vowel opening, and assigned babbles to "phonotactic categories" for infants from 1 to 20 months of age. They found that, of the 11 types of place of articulation recognized by the IPA, 92% of the babbles were one of four types: bilabial, dental-alveolar, velar, and glottal; while 4% were palatal, 3% were uvular, 1% were labiodental, and 0% were all of the remaining 4 types of place of articulation. Of the four infants studied, three initially produced mostly glottal consonants. In addition, they observed variability among and within infants across the age range studied.

The same study also looked at manner of articulation, 9 of which are recognized by the IPA. They found that the infants in the study produced predominantly stops, nasals, and fricatives (91%); and few of the others: semi-vowel (4%), lateral (3%), trill (2%), the remaining three (0%).

Roug et al. (1988) also studied the vowels and found that the vast majority of the vowels were the /a/ in far, the /ae/ in cat, the /e/ in met, the /u/ in but, the /e/ in gate, and the /i/ in bit. An insignificant number of the other vowels recognized by the IPA were observed during this study.

However, the data presented here provide a good starting point for the modeling of babbling in the context of speech production and perception. The main features of the data are the following: (1) The onset of babbling follows a long (relatively uniform in length) period of linguistic experience which may affect vowels, but does not initially affect babbling of consonants. (2) Feedback in the form of auditory or tactile speech is required for the onset of babbling. (3) Future cognitive skill does not seem to affect the time of onset or the content of babbling and late babbling. (4) The babbling repertoire is a small subset of the sounds of the target language. The infant's speech sound repertoire seems to increase rather slowly, only when adult words demand it. Similarly, the babbling repertoire does not depend significantly on the target language (except for some features of vowels). (5) Finally, there is continuity of development with mixing of stages, from the earliest sounds that the infant makes to the acquisition of more than 50 words.

References

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

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

Bedrov, J. A., Chistovich, L. A., & Sheikin, R. L. (1976). Frequency location of the 'center of gravity' of the formants as the useful parameter in vowel perception. *Akust. Zh.*, *24*, 480–486.

Bladon, R. A. V. (1982). Arguments against formants in the auditory representation of speech. In Carlson, R., & Granstrom, B. (Eds.), *The Representation of Speech in the Peripheral Auditory System*, pp. 95–102. Elsevier, Amsterdam.

Blumstein, S. E., & Stevens, K. N. (1980). Perceptual invariance and onset spectra for stop consonants in different vowel environments. *J. Acoust. Soc. Am.*, *67(2)*, 648–662.

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

Borden, G. J. (1980). Use of feedback in established and developing speech. In *Speech and Language: Advances in Basic Research and Practice, Vol 3*, pp. 223–241. Academic Press, Orlando.

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

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 model of motor equivalent reaching and tool use by a multijoint arm. *J. Cognitive Neuroscience*, *5*, 408–435.

Cameron, S. A. (1996). *Self-organizing neural networks for visual navigation and adaptive control*. Ph.D. thesis, Boston University, Boston Massachusetts.

Carpenter, G., Grossberg, S., & Rosen, D. (1991a). Fuzzy ART: Fast stable learning and categorization of analog patterns by an adaptive resonance system. *Neural Networks*, *4*, 759–771.

Carpenter, G. A., Grossberg, S., & Reynolds, J. H. (1991b). Artmap: Supervised real-time learning and classification of nonstationary data by a self-organizing neural network. *Neural Networks*, *4*, 565–588.

Chistovich, L. A. (1985). Central auditory processing of peripheral vowel spectra. *J. Acoust. Soc. Am.*, *77*, 789–805.

Chistovich, L. A., & Lublinskaya, V. V. (1979). The 'center of gravity' effect in vowel

spectra and critical distance between the formants: Psychoacoustical study of the perception of vowel-like stimuli. *Hearing Research*, *1*, 185–195.

Chistovich, L. A., Sheikin, R. L., & Lublinskaya, V. V. (1979). 'Centres of gravity' and spectral peaks as the determinants of vowel quality. In Lindblom, B., & Öhman, S. (Eds.), *Frontiers of Speech Communication Research*, pp. 143–157. Academic Press, London.

Cohen, J. (1989). Applications of an auditory model to speech recognition. *J. Acoust. Soc. Am.*, *85*, 2623–2629.

Daubechies, I. (1988). Orthonormal bases of compactly supported wavelets. *Communications on Pure and Applied Mathematics*, *41*, 909–996.

Daubechies, I. (1992). *Ten Lectures on Wavelets*. SIAM, Philadelphia, PA.

Davis, B., & MacNeilage, P. F. (1990). Acquisition of correct vowel production: A quantitative case study. *J. Speech and Hearing Research*, *33*, 16–27.

de Jong, K. J. (1997). Labiovelar compensation in back vowels. *J. Acoust. Soc. Am.*, *101(4)*, 2221–2233.

Delattre, P. C., Liberman, A. M., Cooper, F. S., & Gerstman, L. J. (1952). An experimental study of the acoustic determinants of vowel colour: Observations on one- and two-formant vowels synthesized from spectrographic patterns. *Word*, *8*, 195–210.

Deng, L., Geisler, C. D., & Greenberg, S. (1988). A composite model of the auditory periphery for the processing of speech. *J. of Phonetics*, *16(1)*, 93.

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

Elbers, L. (1982). Operating principles in repetitive babbling: A cognitive continuity approach. *Cognition*, *12*, 45–63.

Fahey, R. P., Diehl, R. L., & Traunmüller, H. (1996). Perception of back vowels: Effects of varying F1-F0 bark distance. *J. Acoust. Soc. Am.*, *99(4)*, 2350–2357.

Fant, G. (1970). *Acoustic Theory of Speech Production, 2nd Ed.* Mouton, The Hague, The Netherlands.

Flanagan, J. L. (1972). *Speech Analysis, Synthesis, and Perception.* Springer-Verlag, New York.

Flanagan, J. L. (1956). Automatic extraction of formant frequencies from continuous speech. *J. Acoust. Soc. Am.*, *28(1)*, 110–125.

Flanagan, J. L. (1957). Note on the design of terminal-analog speech synthesizers. *J. Acoust. Soc. Am.*, *29*, 306–310.

Fowler, C. A. (1980). Coarticulation and theories of extrinsic timing. *J. Phonetics*, *8*, 113–133.

Ghitza, O. (1988). Temporal non-place information in the auditory nerve firing patterns as a front-end for speech recognition in a noisy environment. *J. of Phonetics*, *16(1)*, 109–124.

Grossmann, A., & Morlet, J. (1984). Decomposition of Hardy functions into square integrable wavelets of constant shape. *SIAM J. Math. Anal*, *15(4)*, 723–736.

Guenther, F. H. (1992). *Neural models of adaptive sensory-motor control for flexible reaching and speaking*. Ph.D. thesis, Boston University, Boston Massachusetts.

Guenther, F. H. (1993). A neural network model of speech acquisition and motor equivalent speech production. Tech. rep. CAS/CNS-TR-93-054, Boston University, Center for Adaptive Systems, Boston.

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

Guenther, F. H. (1994b). Skill acquisition, coarticulation, and rate effects in a neural network model of speech production. *Program of the 127th Meeting of the Acoustical Society of America, J. Acoust. Soc. Am.*, *95(5)*, 2924.

Guenther, F. H. (1995a). A modeling framework for speech motor development and kinematic articulator control. In *Proceedings of the XIIIth International Congress of Phonetic Sciences* Stockholm, Sweden.

Guenther, F. H. (1995b). Speech sound acquisition, coarticulation, and rate effects in a neural network model of speech production. *Psychological Review*, *102(3)*, 594–621.

Guenther, F. H., Hampson, M., & Johnson, D. (1998). A theoretical investigation of reference frames for the planning of speech movements. Psychological Review (In Press).

Guenther, F. H., & Johnson, D. (1995). A computational model using formant space planning of articulator movements for vowel production. *Program of the 129th Meeting of the Acoustical Society of America, J. Acoust. Soc. Am.*, *97(5)*, 3402.

Harshman, R., Ladefoged, P., & Goldstein, L. (1977). Factor analysis of tongue shapes. *J. Acoust. Soc. Am.*, *62*, 693–707.

Heil, P., Langner, G., & Scheich, H. (1992). Processing of FM stimuli in the chick auditory cortex analogue: Evidence of topographic representations and possible mechanisms of rate and directional sensitivity. *J. Comp. Physiol. [A]*, *171*, 583–600.

Heinz, J. M., & Stevens, K. N. (1964). On the derivation of area functions and acoustic spectra from cineradiographic films of speech. *J. Acoust. Soc. Am.*, *36*, 1037–1038.

Holbrook, A., & Fairbanks, G. (1962). Diphthong formants and their movements. *J. Speech and Hearing Research*, *5(1)*, 38–58.

Jackson, M. T. T. (1988). Analysis of tongue positions: Language-specific and cross-linguistic models. *J. Acoust. Soc. Am.*, *84*, 124–143.

Jakobson, R. (1941/68). *Child language, aphasia, and phonological universals.* Mouton, The Hague. Translated by A. R. Keiler.

Johnson, D. (1997). A wavelet-based auditory planning space for production of vowel sounds. In *Proceedings of the Conference on Vision, Recognition, Action: Neural Models of Mind and Machine, May 29-31, 1997*, p. 106 Boston, MA: Boston University.

Johnson, D., & Guenther, F. H. (1995). Acoustic space movement planning in a neural model of motor equivalent vowel production. In *Proceedings of the World Congress on Neural Networks, Vol 1*, pp. 481–484 Washington, D.C.

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

Kaiser, G. (1994). *A Friendly Guide to Wavelets*. Birkhäuser, Boston.

Kent, R. D., & Murray, A. D. (1982). Acoustic features of infant vocalic utterances at 3, 6, and 9 months. *J. Acoust. Soc. Am.*, *72*, 353–365.

Kent, R. D., Osberger, M. J., Netsell, R., & Goldschmidt-Hustedde, C. (1987). Phonetic development in identical twins differing in auditory function. *J. Speech and Hearing Disorders*, *52*, 64–75.

Klatt, D. H. (1980). Software for a cascade/parallel formant synthesizer. *J. Acoust. Soc. Am.*, *67*, 971–995.

Koopmans vanBeinum, F. J., & van der Stelt, J. M. (1986). Early stages in the development of speech movements. In Lindblom, B., & Zetterstrom, R. (Eds.), *Precursors of Early Speech*, pp. 189–204. Stockton Press, New York.

Ladefoged, P. (1964). Physiological parameters of speech. *J. Acoust. Soc. Am.*, *38*, 1037.

Ladefoged, P., & Broadbent, D. E. (1957). Information conveyed by vowels. *J. Acoust. Soc. Am.*, *29(1)*, 98–104.

Lindblom, B. E. F., & Studdert-Kennedy, M. (1967). On the role of formant transitions in vowel recognition. *J. Acoust. Soc. Am.*, *42(4)*, 830–843.

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

Lindblom, B. E. F., & Sundberg, J. E. F. (1971). Acoustical consequences of lip, tongue, jaw, and larynx movement. *J. Acoust. Soc. Am.*, *50*, 1166–1179.

Locke, J. L., & Pearson, D. M. (1990). Linguistic significance of babbling: Evidence from a tracheostomized infant. *J. Child Lang.*, *17*, 1–16.

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

Maeda, S. (1972). Conversion of midsagittal dimensions to vocal tract area function. *Program of the 82nd Meeting of the Acoustical Society of America, J. Acoust. Soc. Am.*, *51*, 89–90.

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

Mallat, S. G. (1989). A theory for multiresolution signal decomposition: The wavelet representation. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, *11(7)*, 674–693.

Mendelson, J. R., & Cynader, M. S. (1985). Sensitivity of cat primary auditory cortex (AI) neurons to the direction and rate of frequency modulation. *Brain Research*, *327*, 331–335.

Mermelstein, P. (1973). Articulatory model for the study of speech production. *J. Acoust. Soc. Am.*, *53*, 1070–1082.

Miller, J. D. (1989). Auditory-perceptual interpretation of the vowel. *J. Acoust. Soc. Am.*, *89(5)*, 2114–2134.

Mitchell, P. R., & Kent, R. D. (1990). Phonetic variation in multisyllable babbling. *Child Language*, *17*, 247–265.

Nossair, Z. B., & Zahorian, S. A. (1991). Dynamic spectral shape features as acoustic correlates for initial stop consonants. *J. Acoust. Soc. Am.*, *89(6)*, 2978–2991.

Öhman, S. E. G. (1966). Coarticulation in VCV utterances: Spectrographic measurements. *J. Acoust. Soc. Am.*, *39*, 151–168.

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

Oller, D. K., Eilers, R. E., Bull, D. H., & Carney, A. E. (1985). Prespeech vocalizations of a deaf infant: A comparison with normal metaphonological development. *J. Speech and Hearing Research*, *28*, 47–63.

Pentland, A. P. (1994). Interpolation using wavelet bases. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, *16(4)*, 410–414.

Perkell, J. S. (1980). Phonetic features and the physiology of speech production. In Butterworth, B. (Ed.), *Language Production, Volume 1: Speech and Talk*, pp. 337–372. Academic Press, New York.

Perkell, J. S., Matthies, M. L., & Svirsky, M. A. (1994). Articulatory evidence for acoustic goals for consonants. *J. Acoust. Soc. Am.*, *96(5)*, 3326.

Perkell, J. S., Matthies, M. L., Svirsky, M. A., & Jordan, M. I. (1993). Trading relations between tongue-body raising and lip rounding in production of the vowel

/u/: A pilot 'motor equivalence' study. *J. Acoust. Soc. Am.*, *93*, 2948–2961.

Peterson, G. E., & Barney, H. L. (1952). Control methods used in a study of the vowels. *J. Acoust. Soc. Am.*, *24(2)*, 175–184.

Plomp, R., Pols, L. C. W., & van de Geer, J. P. (1967). Dimensional analysis of vowel spectra. *J. Acoust. Soc. Am.*, *41*, 707–712.

Pols, L. C. W., van der Kamp, L. J. T., & Plomp, R. (1969). Perceptual and physical space of vowel sounds. *J. Acoust. Soc. Am.*, *46*, 458–467.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., & Flannery, B. P. (1992). *Numerical Recipes in C: The Art of Scientific Computing (Second Edition)*. Cambridge University Press, Cambridge.

Rabiner, L. R., & Schafer, R. W. (1978). *Digital Processing of Speech Signals*. Prentice-Hall, Englewood Cliffs, New Jersey.

Rabiner, L., & Juang, B.-H. (1993). *Fundamentals of Speech Recognition*. Prentice Hall, Englewood Cliffs, New Jersey.

Rioul, O., & Vetterli, M. (1991). Wavelets and signal processing. *IEEE SP Magazine*, *October*, 14–38.

Roug, L., Landberg, I., & Lundberg, L. J. (1988). Phonetic development in early infancy: A study of four swedish children during the first eighteen months of life. *J. Child Lang.*, *16*, 19–40.

Rubin, P., Baer, T., & Mermelstein, P. (1981). An articulatory synthesizer for perceptual research. *J. Acoust. Soc. Am.*, *70*, 321–328.

Sachs, M. B., & Young, E. D. (1979). Encoding of steady state vowels in the auditory-nerve: Representations in terms of discharge rate. *J. Acoust. Soc. Am.*, *66*, 470–479.

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

Savariaux, C., Perrier, P., & Orliaguet, J. P. (1995). Compensation strategies for the perturbation of the rounded vowel [u] using a lip tube: A study of the control space in speech production. *J. Acoust. Soc. Am.*, *98*, 2428–2442.

Schreiner, C. E., & Urbas, J. V. (1986). Representation of amplitude modulation in the auditory cortex of the cat. I. The anterior auditory field (AAF). *Hearing Research*, *21*, 227–241.

Shamma, S. A. (1988). The acoustic features of speech phonemes in a model of the auditory system: Vowels and unvoiced fricatives. *J. of Phonetics*, *16*, 77–91.

Shamma, S. A. (1995). Auditory cortex. In Arbib, M. A. (Ed.), *The Handbook of Brain Theory and Neural Networks*. MIT Press, Cambridge, Massachusetts.

Shamma, S. A., Fleshman, J. W., Wiser, P. R., & Versnel, H. (1993). Organization of response areas in ferret primary auditory cortex. *J. Neurophysiology*, *69(2)*, 367–383.

Shamma, S. A., & Versnel, H. (1995). Ripple analysis in ferret primary auditory cortex. II. Prediction of unit responses to arbitrary spectral profiles. *Auditory Neuroscience*, *1*, 255–270.

Shamma, S. A., Versnel, H., & Kowalski, N. (1995). Ripple analysis in ferret primary

auditory cortex. I. Response characteristics of single units to sinusoidally rippled spectra. *Auditory Neuroscience*, *1*, 233–254.

Smith, B. L., Brown-Sweeney, S., & Stoel-Gammon, C. (1989). A quantitative analysis of reduplicative and variegated babbling. *First Language*, *9*, 175–190.

Stevens, K. N., & Blumstein, S. E. (1978). Invariant cues for place of articulation in stop consonants. *J. Acoust. Soc. Am.*, *64(5)*, 1358–1368.

Stevens, K. N., & House, A. S. (1955). Development of a quantitative description of vowel articulation. *J. Acoust. Soc. Am.*, *27*, 484–493.

Stoel-Gammon, C. (1988). Prelinguistic vocalizations of hearing impaired and normally hearing subjects: A comparison of consonantal inventories. *J. Speech and Hearing Disorders*, *53*, 302–315.

Stoel-Gammon, C. (1989). Prespeech and early speech development of two late talkers. *First Language*, *9*, 207–224.

Stoel-Gammon, C., & Otomo, K. (1986). Babbling development of hearing impaired and normally hearing subjects. *J. Speech and Hearing Disorders*, *51*, 33–41.

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

Strange, W. (1989). Evolving theories of vowel perception. *J. Acoust. Soc. Am.*, *85(5)*, 2081–2087.

Sussman, H. M., Fruchter, D., Hilbert, J., & Sirosh, J. (1998). The orderly output constraint: A functional role for highly correlated, linearly related components in the speech signal. Behavioral and Brain Sciences (In Press).

Sussman, H. M., McCaffrey, H. A., & Matthews, S. A. (1991). An investigation of locus equations as a source of relational invariance for stop place categorization. *J. Acoust. Soc. Am.*, *90(3)*, 1309–1325.

Syrdal, A. K., & Gopal, H. S. (1986). A perceptual model of vowel recognition based on the auditory representation of American English vowels. *J. Acoust. Soc. Am.*, *79*, 1086–1100.

Traunmüller, H. (1982). Perception of timbre: Evidence for spectral resolution bandwidth different from critical band?. In Carlson, R., & Granstrom, B. (Eds.), *The Representation of Speech in the Peripheral Auditory System*. Elsevier, New York.

Traunmüller, H. (1984). Articulatory and perceptual factors controlling the age- and sex-conditioned variablility in formant frequencies of vowels. *Speech Communication*, *3*, 49–61.

Vetterli, M., & Kovačević, J. (1995). *Wavelets and Subband Coding*. Prentice Hall Inc., Englewood Cliffs, New Jersey.

Wang, K., & Shamma, S. A. (1994a). Modeling the auditory functions in the primary cortex. *Optical Engineering*, *33(7)*, 2143–2148.

Wang, K., & Shamma, S. A. (1994b). Self-normalization and noise robustness in early auditory processing. *IEEE Trans. on Speech and Audio Processing*, *2(?)*, 421–435.

Wang, K., & Shamma, S. A. (1995). Spectral shape analysis in the central auditory system. *IEEE Trans. on Speech and Audio Processing*, *3(5)*, 382–395.

Yang, X., Wang, K., & Shamma, S. (1992). Auditory representations of acoustic signals. *IEEE Trans. on Info. Theory*, *38*, 824–839.

Zahorian, S. A., & Jagharghi, A. J. (1993). Spectral-shape features versus formants as acoustic correlates for vowels. *J. Acoust. Soc. Am.*, *94(4)*, 1966–1982.

Zahorian, S. A., & Rothenberg, M. (1981). Principal-components analysis for low-redundancy encoding of speech spectra. *J. Acoust. Soc. Am.*, *69*, 832–845.

Zwicker, E. (1961). Subdivision of the audible frequency range into critical bands (frequenzgruppen). *J. Acoust. Soc. Am.*, *33*, 248.

Zwicker, E., & Terhardt, E. (1980). Analytical expressions for critical-band rate and critical bandwidth as a function of frequency. *J. Acoust. Soc. Am.*, *68*, 1523–1525.

Vita

Dave Johnson was born at Fort Riley, Kansas, April 30, 1957, and grew up in Kansas, Indiana, Iowa, Kentucky, and Illinois. He completed high school in 1975, then completed his B.S. in Engineering Physics at University of Illinois, Urbana-Champaign, in May, 1979. He moved to San Jose, California, and worked full-time for Hewlett-Packard, then later for Schlumberger Ltd. During this time he also pursued and received his M.S. in Applied Mathematics from Santa Clara University in December, 1983. He then transferred to Austin, Texas, where he worked in several software engineering positions. In August 1993, he returned to graduate school to study at Boston University in the Department of Cognitive and Neural Systems.

Publications

Frank H. Guenther, Michelle Hampson, and Dave Johnson (1998), A theoretical investigation of reference frames for the planning of speech movements. *Psychological Review.* In press.

Dave Johnson (1997). A wavelet-based auditory planning space for production of vowel sounds. *Proceedings of the Conference on Vision, Recognition, Action: Neural Models of Mind and Machine.* May 29-31, 1997. Boston, MA: Boston University, 106.

Frank H. Guenther and Dave Johnson (1995), A computational model using formant space planning of articulator movements for vowel production. *Journal of*

the Acoustical Society of America, 97(5), 3402.

Dave Johnson and Frank H. Guenther (1995), Acoustic space movement planning in a neural model of motor equivalent vowel production. *Proceedings of the World Congress on Neural Networks, Washington, D.C.*, Vol 1, 481-484, Mahwah, NJ: Lawrence Erlbaum Associates, Inc., Publishers.