UNIVERSITY OF CALIFORNIA, SAN DIEGO

Visualizing Speech with a Recurrent Neural Network

Trained on Human Acoustic-Articulatory Data

A dissertation submitted in partial satisfaction of the

requirements for the degree Doctor of Philosophy

in Cognitive Science

by

Jay T. Moody

Committee in charge:

Professor Jeffrey L. Elman, Co-Chair
Professor Maureen Stone, Co-Chair
Professor John Batali
Professor Garrison W. Cottrell
Professor Javier R. Movellan
Professor Terrence J. Sejnowski

1999
The dissertation of Jay T. Moody is approved, and it is acceptable in quality and form for publication on microfilm:

________________________________________

________________________________________

________________________________________

________________________________________

________________________________________

Co-Chair

Co-Chair

University of California, San Diego

1999
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Signature Page</td>
<td>iii</td>
</tr>
<tr>
<td>Table of Contents</td>
<td>iv</td>
</tr>
<tr>
<td>List of Figures</td>
<td>vi</td>
</tr>
<tr>
<td>List of Tables</td>
<td>ix</td>
</tr>
<tr>
<td>Vita and Publications</td>
<td>x</td>
</tr>
<tr>
<td>Abstract</td>
<td>xi</td>
</tr>
<tr>
<td>Chapter 1: Introduction</td>
<td>1</td>
</tr>
<tr>
<td>Chapter 2: Previous Approaches to the Vocal Tract Inverse Problem</td>
<td>5</td>
</tr>
<tr>
<td>The Analytical Approach</td>
<td>5</td>
</tr>
<tr>
<td>The Empirical Approach</td>
<td>8</td>
</tr>
<tr>
<td>Summary: Where more research is needed</td>
<td>25</td>
</tr>
<tr>
<td>Chapter 3: Learning the Dynamics of Voiced Stops in a Schwa Context</td>
<td>27</td>
</tr>
<tr>
<td>Part One: Principal Component Analysis of Lip and Tongue Images</td>
<td>27</td>
</tr>
<tr>
<td>Methods</td>
<td>27</td>
</tr>
<tr>
<td>Results and Discussion</td>
<td>37</td>
</tr>
<tr>
<td>Part Two: Training a Neural Network to Estimate Articulation from</td>
<td>53</td>
</tr>
<tr>
<td>Acoustics</td>
<td>53</td>
</tr>
<tr>
<td>Methods</td>
<td>56</td>
</tr>
<tr>
<td>Chapter 4: Improving the Method — Long Windows and Painted Faces</td>
<td>65</td>
</tr>
<tr>
<td>Methods</td>
<td>65</td>
</tr>
<tr>
<td>PCA Results and Discussion</td>
<td>69</td>
</tr>
<tr>
<td>Neural Network Results and Discussion</td>
<td>75</td>
</tr>
<tr>
<td>Chapter 5: Generating Speechreadable Vocal Tract Images from Acoustic Input</td>
<td>85</td>
</tr>
</tbody>
</table>
Chapter 6: Experimental Conclusions and Theoretical Implications — A New Hypothesis for Human Speech Perception

Summary of Experimental Findings
Implications for Theories of Human Speech Perception
A New Hypothesis
Behavioral and Neuropsychological Evidence for the Hypothesis
Experimental Predictions of the Hypothesis

References
LIST OF FIGURES

Figure 3.1: The HATS system for ultrasound recording of the vocal tract................. 28
Figure 3.2: Spectrally-weighted bark bins ................................................................. 32
Figure 3.3: The effect of various steps in the audio processing procedure ............... 33
Figure 3.4: Sample images from the production of /bɔ/, /dɔ/, and /ɡɔ/...................... 38
Figure 3.5: The average lip and average tongue images ........................................... 38
Figure 3.6: The first seven principal components of the lips and tongue shown

with maximal contrast.................................................................................................. 39
Figure 3.7: The first seven principal components of the lips and tongue

displayed to indicate relative importance................................................................. 39
Figure 3.8: The percent variance explained by the first 20 PCs of the lip and

tongue images........................................................................................................... 40
Figure 3.9: Reconstructions of lip images from various numbers of PCs ................. 41
Figure 3.10: Reconstructions of tongue images from various numbers of PCs.......... 42
Figure 3.11: Lip images reconstructed by setting the loading on each principal

component to its maximum and minimum values .................................................... 43
Figure 3.12: Tongue images reconstructed by setting the loading on each

principal component to its maximum and minimum values...................................... 43
Figure 3.13: Variation in lip PC loads over time for three sample utterances ......... 45
Figure 3.14: Variability in the production of /bɔ/....................................................... 46
Figure 3.15: Differences in lip motion during the opening and closing of /bɔb/......... 47
Figure 3.16: Variation in tongue PC loads over time for three sample utterances. .... 48
Figure 3.17: Tongue load trajectories in PC1-PC2 space for all nine utterances........ 49
Figure 3.18: Lip load trajectories in PC1-PC2 space for the /bɔbɔbɔ/ utterances

(a), the /dɔdɔdɔ/ utterances (b), and the /ɡɔɡɔɡɔ/ utterances (c)............................... 50
Figure 3.19: Neural network architectures ................................................................. 53
Figure 3.20: Performance of the best FFN on novel test utterances ......................... 59
Figure 3.21: Performance of the best SRN on novel test utterances ......................... 60
Figure 3.22: Sample articulatory estimates for /bøb/ by the best FFN and best SRN ................................................................. 61
Figure 3.23: Lip PC2 targets (solid) and estimations by the best SRN (dashed)
on the /dødødø/ utterances ........................................................................ 63
Figure 4.1: Sample images from the three utterance types .......................................... 70
Figure 4.2: The average lip and average tongue images ............................................. 70
Figure 4.3: The first seven principal components of the lips and tongue scaled in intensity to indicate relative importance ............................................ 70
Figure 4.4: The percent variance explained by the first 20 PCs of the lip and
tongue images ............................................................................................ 71
Figure 4.5: Variation in lip PC loads over time for three sample utterances ................. 72
Figure 4.6: Differences in lip motion during the opening and closing of /øba/ .............. 73
Figure 4.7: Variability in the production of /øda/ ..................................................... 73
Figure 4.8: Variation in tongue PC loads over time for three sample utterances ........... 74
Figure 4.9: Tongue image locations in PC1-PC3 space for the 10 /øda øda øda/
and /øga øga øga/ utterances. Coarticulation of the schwa is manifest in the lack of overlap between the two utterance types ............ 75
Figure 4.10: PC1 and PC3 estimation error for the /øda øda øda/ and /øga øga
øga/ utterances by the SRNs (x) and the FFNs (o) trained on the
tongue ........................................................................................................... 78
Figure 4.11: Lip images from /øba/ for the best FFN and the best SRN ....................... 79
Figure 4.12: Tongue images from /øda/ for the best FFN and the best SRN ............... 80
Figure 4.13: Tongue images from /øba/ for the best FFN and the best SRN ............... 81
Figure 4.14: Lip images from the sixth and tenth repetitions of the /əʊəʊəʊəʊə/ utterance. Lip aperture in the sixth repetition was unusually small. ....................................................................................................... 83

Figure 4.15: Failure to estimate exceptionally small lip aperture in the sixth repetition of /əʊəʊəʊəʊə/........................................................................................................ 83

Figure 5.1: The effect of low-pass filtering at 100 Hz as shown by the spectrogram of the utterance “say five four seven.” .......................... 88

Figure 5.2: The 24 most distinct lip and tongue images from the “numbers” speech corpus .................................................................................................................. 89

Figure 5.3: The average lip and average tongue images .......................................................................................................................... 90

Figure 5.4: The first eight principal components of the lips and tongue scaled in intensity to indicate relative importance .................................................................................. 90

Figure 5.5: The percent variance explained by the first 20 PCs of the lip and tongue images.......................................................................................... 90

Figure 5.6: Percent correct recognition of spoken digits when no audio signal was available .......................................................................................................................... 93

Figure 5.7: Percent correct recognition of spoken digits under conditions of simulated hearing impairment........................................................................ 94
LIST OF TABLES

Table 2.1: A summary of representative vocal tract inverse problem studies which have used the empirical approach................................................................. 9

Table 3.1: Average correlation coefficients (with s.d.) between target and network output trajectories for the untrained test utterances of the nine networks of each type........................................................................ 58

Table 3.2: Average root mean square error (with s.d.) on the untrained test utterances for the nine networks of each type.................................................. 58

Table 4.1: Average correlation coefficients (with s.d.) between target and network output trajectories for the most critical PCs of each test utterance (see text). ........................................................................ 76

Table 4.2: Average root mean square error (with s.d.) on the test utterances. ................................................................................................................. 77

Table 5.1: Digit sequences used in the speech corpus .................................................................................................................. 86

Table 5.2: Average root mean square error (with s.d.) and correlation coefficients (with s.d.) between target and network output trajectories for all eight PCs of each composite network.................. 91

Table 5.3: Percent correct spoken digit recognition (with s.d.) for familiar vs. novel utterances..................................................................................... 95

Table 5.4: Confusion matrices for the audio, visual, and audio-visual conditions in which only original video was presented........................................ 96

Table 5.5: Confusion matrices for the visual and audio-visual conditions using network output video........................................................................... 97
VITA

August 5, 1965  Born, Abilene, Texas
1987  B.A.  Austin College, Sherman, Texas
       Chemistry
1992  M.A.  The Ohio State University
       Japanese Linguistics
1999  Ph.D.  University of California, San Diego
       Cognitive Science

PUBLICATIONS


A method is presented for converting acoustic speech data into a visual representation of articulation using a recurrent neural network. Target values for the network were created by finding the principal components of sets of video frames representing the face and tongue of the speaker. Midsagittal images of the tongue were captured using ultrasound. When trained on a corpus of voiced stops, the network estimated the magnitude of the principal component loads with an average accuracy of 14% of the range of variation. When the load estimates were converted into moving images, the appropriate bilabial, alveolar, and velar gestures were clearly apparent. Feedforward networks trained on the same task were unable to distinguish the consonants, indicating
more inconsistency in the static acoustic-articulatory mapping than in the mapping between dynamic acoustic and articulatory trajectories. Because the networks averaged over disparate targets associated with similar inputs, they tended to under-represent articulatory aberrations, in a sense “cleaning up” articulatory variability. In a direct test of the usefulness of the estimated articulatory images, human subjects were shown, under conditions of simulated hearing impairment, images of moving lips produced by a network trained on a corpus of spoken digits. Performance rose from less than 50% to over 80% correct recognition when the acoustically-driven network output movies were displayed. In light of this performance, and inspired by the rough similarities between artificial and biological neural networks, a new hypothesis for human speech perception is formulated in which articulatory representations play an important role. The hypothesis differs from established motor theories of speech perception in that gestural representations are not taken to be innate or essential, but instead learned through experience and developed in response to low-level neurophysiological pressures rather than to behavioral-level needs.
CHAPTER 1: INTRODUCTION

Speech sounds are generated by passing an excitation signal through the vocal tract. The vocal cavities act as an acoustic tube, modifying the spectrum of the excitation signal to create recognizable speech sounds. This forward transformation — from the shape of the vocal tract to the acoustic characteristics of a sound — is well understood (Fant, 1960; Flanagan, 1972). That is, given a description of the shape of the vocal tract, the resulting sound can be accurately estimated. The inverse transformation — from speech acoustics to the vocal tract shapes that produced them — is not well understood. The problem of estimating vocal tract shapes from speech sounds is known as the vocal tract inverse problem.

A solution to the vocal tract inverse problem would have a variety of practical, clinical, and theoretical applications. For example, the ability to accurately estimate lip movements from sounds might lead to a “speechreadable” display for persons with hearing loss, at least in situations in which the audio signal could be channeled to a computer with a video display (such as on the telephone or in a specially equipped classroom). And the ability to estimate tongue positions might lead to a visual feedback system to help persons undergoing speech therapy. A solution to the vocal tract inverse problem might also improve automatic speech recognition (Rose, Schroeter & Sondhi, 1996; Zlokarnik, 1995) and speech coding (Flanagan, Ishizaka & Shipley, 1980; Schroeter & Sondhi, 1994), by allowing the incorporation of constraints from the articulatory domain. Further, the issue of whether or not the inverse problem is solvable has important implications for theories which claim that articulatory gestures are recovered by humans as a part of the speech perception process (Fowler, 1986; Liberman, Cooper, Shankweiler & Studdert-Kennedy, 1967; Liberman & Mattingly, 1985).
Although the vocal tract inverse problem can be simply stated, it involves many subtleties. Consider, for example, the speech of an accomplished ventriloquist. Does the fact that she can produce clear speech with an unusual vocal tract shape imply that the inverse problem cannot be solved? If one takes a high level view of speech, focusing on the perceived meaningful sounds as the important units, it would seem indeed that the vocal tract inverse problem is *prima facie* insoluble because the ventriloquist is able to produce the “same” sounds with different vocal tract shapes. From a lower level view, however, we notice that the acoustic signal may differ even if the perceived phonemes do not, and that these differences might provide a basis upon which the vocal tract shapes could be distinguished. In this sense, the *uniqueness* of the vocal tract inverse mapping (the notion of whether a particular sound maps uniquely onto a single vocal tract shape) depends on the level of precision one adopts in defining whether or not two sounds are or are not the “same.”

Another subtlety of the vocal tract inverse problem involves the issue of time. Consider stop consonants, for example, in which there is a brief period of silence as the vocal tract is completely occluded. At the moment of silence of the stop, the vocal tract shape must be indeterminable from acoustics alone because silence does not imply any particular vocal tract configuration. But this is true only if we assume a static version of the inverse problem in which it is asked whether individual acoustic segments can be mapped to static vocal tract shapes without regard for the context in which the acoustic-articulatory pair is found. A dynamic version of the problem would ask instead whether acoustic trajectories could be used to estimate articulatory trajectories. When temporal context is considered, the estimation of stop consonant articulation may well become feasible.

It is thus inappropriate to state in general terms whether the vocal tract inverse mapping is or is not unique. The question of real interest is much less elegant than the
simple statement of the problem — “can articulation be estimated from acoustics?” — would imply. The question of real interest is: can articulation be estimated from acoustics for enough sounds and with enough accuracy to be useful in speech recognition? To see that use could be made of an inverse mapping that is less than perfectly unique, consider a (human or artificial) speech recognizer that, by hypothesis, is able to improve its recognition accuracy by incorporating inferred knowledge about articulation. If the articulatory inference mechanism is constructed such that it produces only estimates of typical articulations, then when confronted with speech produced by an atypical articulation (consider again the ventriloquist), the estimates will be inaccurate in real terms but might nevertheless help the recognizer understand the speech (depending, of course, on how the articulatory inference mechanism responds to whatever acoustic differences are caused by the differences in articulation). An acoustically-driven speechreadable display might similarly produce articulatory movements that are speechreadable but different from actually executed atypical movements.

The study presented here addresses this less elegant version of the vocal tract inverse problem. The study is concerned not with determining whether the mapping from acoustics to articulation is strictly unique, but whether it is consistent enough under normal speaking conditions for articulation to be estimated with sufficient accuracy for the recovered gestures to be identified. The study employs a novel technique for representing articulatory states which is based on the application of principal component analysis to raw video images of the lips and ultrasound images of the tongue. The study incorporates dynamics into the acoustic-articulatory mapping through the use of a recurrent neural network, and is able to demonstrate the advantage of the dynamical mapping procedure over a static one on a corpus of temporally challenging stop consonants. A corpus of real words including 17 different phonemes is also studied, and the estimated lip images are tested for their speechreadability under conditions of
simulated hearing impairment. The study also takes a keen interest in the question of whether humans might perform articulatory inference in the process of speech perception. Although the methods used are for the most part not biologically realistic, the method used for acoustic analysis mimics in a very general sense the activity of the human cochlea, and the artificial neural networks used to perform the acoustic-articulatory mapping can be thought of as highly abstract models of the gross behavior of human neural systems. The process of applying such models to the vocal tract inverse problem has brought insight into the potential role of articulatory inference in human speech perception. Before detailing the approach taken in this study, however, let us look at the approaches others have taken to the vocal tract inverse problem.
CHAPTER 2: PREVIOUS APPROACHES TO THE VOCAL TRACT INVERSE PROBLEM

The approaches that have been taken with regard to the vocal tract inverse problem can be classified into two basic types, analytical and empirical. In the analytical approach an attempt is made to (analytically) invert mathematical descriptions of sound wave propagation in ideal acoustic tubes. Studies using this approach have contributed greatly to our understanding of the vocal tract inverse problem, but have not in the end produced practical inversion methods. The empirical approach, by contrast, seeks a numerical mapping between a large number of empirically derived acoustic-articulatory data points which are either measured from actual human speech (as in the current study) or generated from a speech production model. This section reviews each of these types of methods, starting with the analytical method which was tried, and abandoned, first.

The Analytical Approach

Some of the earliest studies of the vocal tract inverse problem argued that a large class of vocal tract shapes could not be determined from the measurement of normal speech acoustics. Based on the physics of wave propagation in an ideal acoustic tube, Mermelstein (1967) and Schroeder (1967) showed that the shape of the tube could be uniquely determined from its resonance frequencies (in the speech domain called formant frequencies) only if the resonance frequencies from two different sets of boundary conditions were known. In particular, the resonance frequencies in the condition of “one end open and one end closed” determined the odd Fourier coefficients of the log area function, and the resonance frequencies in the condition of “both ends closed (or open)”
determined the even coefficients. ¹ As the glottal end of the human vocal tract represents a closed boundary condition, the implication of this result was that in order to obtain a complete description of the area function one would have to simultaneously measure the resonance frequencies of the vocal tract with the lip end open and closed — a seemingly impossible task.

Mermelstein (1967) showed, however, that the simultaneous resonance frequencies of both of the boundary conditions (lips open and closed) could be obtained by measuring the impulse response at the lips. The impulse response was obtained by measuring the pressure developed at the lips in response to a unit impulse (ideally an acoustic pulse of infinitely large amplitude and infinitely short duration, but in practice a very brief acoustic pulse) delivered from an external source into the vocal tract. In order to make such a measurement the speaker's lips must be coupled with a special seal to an impedance-measuring tube and the speaker must articulate without phonation. The computational procedure was improved by Paige & Zue (1970), Sondhi & Gopinath (1970), and Sondhi & Resnik (1983). However, even though a vocal tract inverse problem of sorts can be solved with this method (in the sense that area functions can be obtained from acoustical measurements), it is of limited practical value since it cannot be applied to the acoustics of normal speech.

Other studies continued the search for an analytical solution to the inverse problem based on natural speech acoustics. Atal (1970), Wakita (1973), and Wakita & Gray (1975) developed a method based on inverse filtering which showed that, under certain assumptions, area functions could be calculated as long as the formant frequencies

¹The log area function represents the vocal tract shape as the natural logarithm of the cross-sectional area at points taken along the long axis of the tract, i.e., from the glottis to the lips. The Fourier coefficients of this function describe the sine wave components of the shape at integral multiples of frequency.
and their bandwidths were known. The assumptions implied by treating the vocal tract as an ideal acoustic tube included a host of approximations, however. Specifically, it was assumed:

1) that the walls of the vocal tract are rigid and dry so that there is no energy loss,
2) that the boundary conditions are ideal, so that in the open condition that there is no radiation load and in the closed condition (including the glottis) there is perfect energy loss at all frequencies,
3) that the vocal tract has a rate of change of cross-sectional area with distance along the tract that is small, i.e., that there are no step-like discontinuities,
4) that the main axis of the vocal tract may be straightened out without changing its resonances,
5) that pressure and velocity are constant in a plane perpendicular to the straightened out axis of the tract, and
6) that the linear wave equation is valid.

The approximations introduced by each of these assumptions are expected to introduce some error into articulatory estimations. Of particular concern are the facts that, 1) energy losses due to vocal tract wall vibration have appreciable effects on formant frequencies and bandwidths and boundary conditions are far from ideal (Wakita, 1979); 2) step-like discontinuities do exist near the tongue tip in many vocal tract configurations; 3) although it can be demonstrated that a tube of uniform cross-section may be straightened without changing its resonances (Sondhi, 1986), cross-sectional shapes of human vocal tracts are far from uniform (Stone, 1991); and 4) the linear wave equation loses validity at extremely narrow constrictions, thus reducing its applicability for consonantal sounds (Flanagan, 1972).
In addition to these questionable assumptions, there are a number of practical issues that further limit the accuracy and applicability of analytical methods. First, the spatial resolution of the estimated area function is limited by the number of measurable formants, and by the accuracy with which their frequencies and bandwidths can be measured. Reliable measurement of formants can in practice be quite difficult, especially for formants four and higher. Second, the methods ignore excitation source characteristics which can interact with the vocal tract shape to affect formant amplitudes and bandwidths (Larar, Alsaka & Childers, 1985; Schroeter & Sondhi, 1994). Third, the calculations of vocal tract areas depend on knowledge of the vocal tract length, but tract length varies considerably between speakers and the effective length varies even within a single speaker as the lips protrude and flatten. Fourth, because the methods do not include a side branching tube, they are inappropriate for nasal sounds (whether consonants or vowels). Because of these several limitations, the analytical approach was largely abandoned in the late 1970s in favor of the empirical approach.

The Empirical Approach

The empirical approach encompasses methods known in the inverse problem literature as codebook, sorting, analysis-by-synthesis, neural network, and other statistical mapping methods. A large number of studies fall into the empirical category. Instead of reviewing each study individually, a synthesis is provided in the form of Table 2.1. This table lists several representative studies of the vocal tract inverse problem and highlights six issues of particular relevance to the current study: 1) what type of mapping methods have been used, 2) what sources have been used for the generation of acoustic-articulatory data, 3) how accurate articulatory estimations have been, 4) whether and how dynamic aspects of the mapping have been incorporated, 5) how acoustic signals have been analyzed, and 6) what range of speech sounds have been studied. By reviewing
Table 2.1: A summary of representative vocal tract inverse problem studies which have used the empirical approach. Terms in the table are defined in the text. Empty cells indicate that information for that cell could not be determined.

<table>
<thead>
<tr>
<th>Study</th>
<th>(1) Mapping Method</th>
<th>(2) Data Source</th>
<th>(3) Estimation Error</th>
<th>(4) Dynamics</th>
<th>(5) Acoustics</th>
<th>(6) Sounds Studied</th>
</tr>
</thead>
<tbody>
<tr>
<td>Atal et al., 1978</td>
<td>DB (Lookup)</td>
<td>Model</td>
<td>_</td>
<td>None</td>
<td>Formant freq</td>
<td>Vowels</td>
</tr>
<tr>
<td>Ladefoged et al., 1978</td>
<td>DB (Regres)</td>
<td>Real (cinerad)</td>
<td>~13%</td>
<td>None</td>
<td>Formant freq</td>
<td>Vowels</td>
</tr>
<tr>
<td>Flanagan et al., 1980</td>
<td>ABS</td>
<td>Model</td>
<td>Casual</td>
<td>None</td>
<td>PS</td>
<td>Vowels</td>
</tr>
<tr>
<td>Levinson &amp; Schmidt, 1983</td>
<td>ABS</td>
<td>Model</td>
<td>Casual</td>
<td>“Initialize with prev”</td>
<td>LPC spectrum</td>
<td>Vowels</td>
</tr>
<tr>
<td>Schroeter &amp; Sondhi, 1986</td>
<td>ABS</td>
<td>Model</td>
<td>_</td>
<td>None</td>
<td>PS</td>
<td>Vowels</td>
</tr>
<tr>
<td>Shirai &amp; Kobayashi, 1986</td>
<td>ABS</td>
<td>Model</td>
<td>_</td>
<td>Kalman</td>
<td>Several</td>
<td>Vowels</td>
</tr>
<tr>
<td>Atal &amp; Rioul, 1989</td>
<td>DB (NN)</td>
<td>Model</td>
<td>_</td>
<td>None</td>
<td>LPC coefficients</td>
<td>Vowels</td>
</tr>
<tr>
<td>Rahim &amp; Goodyear, 1990</td>
<td>DB (NN)</td>
<td>Model</td>
<td>_</td>
<td>None</td>
<td>PS</td>
<td>Vowels, nasals</td>
</tr>
<tr>
<td>Strube et al, 1989-91</td>
<td>DB (NN)</td>
<td>Model</td>
<td>_</td>
<td>Smoothing, Kalman, DP</td>
<td>LPC coef / PS (Bark)</td>
<td>–</td>
</tr>
<tr>
<td>Rahim et al., 1991</td>
<td>DB (NN)</td>
<td>Model</td>
<td>_</td>
<td>Smoothing and DP</td>
<td>Cepstral coefficients</td>
<td>Vowels, nasals, glides</td>
</tr>
<tr>
<td>Shirai &amp; Kobayashi, 1991</td>
<td>DB (NN)</td>
<td>Model</td>
<td>_</td>
<td>None</td>
<td>Cepstral coefficients</td>
<td>Vowels</td>
</tr>
<tr>
<td>Soquet et al, 1991</td>
<td>DB (NN)</td>
<td>Model</td>
<td>_</td>
<td>_</td>
<td>_</td>
<td>Vowels</td>
</tr>
<tr>
<td>Boë et al., 1992</td>
<td>DB (Lookup)</td>
<td>Model</td>
<td>_</td>
<td>None</td>
<td>Formant freq, bdwdth, amp</td>
<td>Vowels</td>
</tr>
<tr>
<td>Papçun et al., 1992</td>
<td>DB (NN)</td>
<td>Real (x-ray µ)</td>
<td>8-14%</td>
<td>Wide window</td>
<td>PS (Bark)</td>
<td>Stops in schwa context</td>
</tr>
<tr>
<td>Rahim et al., 1993</td>
<td>DB (NN)</td>
<td>Model</td>
<td>_</td>
<td>DP</td>
<td>Cepstral coefficients</td>
<td>–</td>
</tr>
<tr>
<td>Schroeter &amp; Sondhi, 1992, 1994</td>
<td>DB (Lookup)</td>
<td>Model</td>
<td>Casual</td>
<td>DP</td>
<td>LPC coefficients</td>
<td>Broad range</td>
</tr>
<tr>
<td>Zacks &amp; Thomas, 1994</td>
<td>DB (NN)</td>
<td>Real (x-ray µ)</td>
<td>4%</td>
<td>Wide window</td>
<td>PS (Bark)</td>
<td>Vowels in /kVd/</td>
</tr>
<tr>
<td>Zlokarnik, 1995</td>
<td>DB (NN)</td>
<td>Real (EMMA)</td>
<td>_</td>
<td>_</td>
<td>_</td>
<td>Broad range in V₁CV₂</td>
</tr>
<tr>
<td>Lavagetto, 1995, 1997</td>
<td>DB (NN)</td>
<td>Real (video)</td>
<td>_</td>
<td>Wide window</td>
<td>Cepstral coefficients</td>
<td>Broad range</td>
</tr>
<tr>
<td>Hogden et al., 1996</td>
<td>DB (Lookup)</td>
<td>Real (EMMA)</td>
<td>9-15%</td>
<td>None</td>
<td>PS (smoothed)</td>
<td>Vowels in /gVVg/</td>
</tr>
</tbody>
</table>
these six issues we can see where attention has been focused over the years and where more attention needs to be directed.

**Mapping Methods**

Column 1 of Table 2.1 lists the methods that have been used to solve the acoustic–articulatory mapping in each study. Studies denoted DB begin by amassing a database of acoustic-articulatory pairs; they then find articulatory configurations for a given sound using some sort of mapping technique. These techniques include (1) training a neural network to map acoustic to articulatory vectors (NN), (2) quantizing and sorting the database so that the articulatory configuration corresponding to a given sound can be found by table look-up (Lookup), (3) statistical regression techniques (Regres). Studies denoted ABS use the analysis-by-synthesis method in which the vector pairs are not pre-computed. Instead, input speech is compared to speech synthesized on the basis of a model of human speech production. The parameters of the speech production model are then iteratively adjusted until the synthesized speech is similar to the input speech. The final conditions found for the speech production model are then assumed to be similar to the vocal tract conditions that produced the input speech.

**Acoustic and Articulatory Data Sources**

**Speech Production Models**

Column 2 of Table 2.1 lists the sources that have been used for generating the articulatory–acoustic data. Model indicates that the data were obtained from a speech production model, such as the models of Maeda (1979, 1988), Mermelstein (1973), or models developed in each particular study. Most models describe the midsagittal view of the vocal tract by specifying the positions of various articulators (for example, in Mermelstein's model: jaw, tongue body, tongue tip, lips, velum and hyoid) or the location
and degree of various openings (e.g., maximum constriction location, maximum constriction degree, lip aperture area, etc.). This sagittal representation is then converted to an area function, which is used to calculate the transfer function of the vocal tract (a description of how much energy is passed at each frequency) based on a knowledge of wave propagation in acoustic tubes.

Speech production models make some of the same simplifying assumptions made under the analytical approach. For example, they typically assume the validity of the linear wave equation, which holds only as long as there are no extremely narrow constrictions. The models thus tend to have difficulty with consonantal sounds. The models also make simplifying assumptions about losses in the vocal tract due to lip radiation, viscosity, heat conduction, wall vibration, and glottal effects. While such losses are accounted for in some studies (Atal, Chang, Mathews & Tukey, 1978; Boë, Perrier & Bailly, 1992; Rahim, Kleijn, Schroeter & Goodyear, 1991), they are simply ignored in others.

Another complication of using a model is that it is difficult to limit the vocal tract shapes obtainable by the model to all and only those that are physiologically possible. Studies have either sampled all articulatory parameters along their entire allowed ranges (the “random sampling method”), which produces many combinations of parameters which are unrealistic, or they have sampled realistic extreme vocal tract shapes and all of the paths between the chosen extremes (“the root-shape interpolation method”), which tends to omit many plausible shapes (Larar, Schroeter & Sondhi, 1988; Schroeter & Sondhi, 1994).

Given these several limitations we might ask how these model-based studies have performed. Unfortunately, the results have seldom been validated against actual articulatory measurements (see below). What we do know is that the choice of speech
production model may drastically affect the findings of vocal tract inverse problem studies (Hogden et al., 1996).

Measurement of Human Articulation

Some studies avoid the pitfalls of production models by making simultaneous acoustic and articulatory measurements of real human vocal tracts (indicated Real in Table 2.1). This can be a very difficult task, however, as many articulators are difficult to access without interfering with normal speech. In addition, the lips and tongue are very flexible structures with enormous numbers of degrees of freedom in their movements. And these structures take on even greater complexity by interacting with the structures surrounding them. The tongue, for example, may alter its shape by bracing itself against the palate or teeth, and the lips may use resistance from the teeth, each other, or even air pressure that is built up in the oral cavity. And both the tongue and the lips are partially supported by the mandible, which is itself a moving structure. All of these characteristics combine to make articulatory movements difficult to measure and describe — a fact that at least partially explains the common use of articulatory models in spite of their known limitations (Jospa & Soquet, 1994). Several methods of articulatory measurement have been developed, however, each with its own features and failures. The methods that have been used in vocal tract inverse problem studies include the following.

Cineradiography

Cineradiography (cinerad in Table 2.1) is full motion x-ray. Compared to other methods, cineradiography provides very comprehensive information about the vocal tract. There are some problems associated with projecting the entire transverse dimension onto a single plane, however. For example, the most elevated regions of the tongue create the line usually taken as the tongue surface, but because of arching and grooving in the
transverse plane, this line may or may not represent the midsagittal contour (Heike, 1979; Stone, 1990). Also, the tongue surface line is often obscured by the teeth. Of course the main disadvantage of cineradiography, and the reason that it is no longer used for research purposes, is the high level of radiation exposure it entails.

**X-ray Microbeam**

X-ray microbeam (indicated \(X\)-ray \(\mu\) in Table 2.1), reduces radiation exposure to safe levels by using a 1-mm diameter x-ray beam to track movements of gold pellets attached to different articulators with dental adhesive (Abbs, Nadler & Fujimura, 1988; Kiritani, Itoh & Fujimura, 1975). Because this method provides coordinates of points on individual articulators, the data provide a representation of articulatory activity that is sparse but relatively easy to analyze. Limitations of x-ray microbeam include its high cost and limited availability, and the fact that pellets cannot be placed on the rear portions of the tongue because of the gag reflex. Also, there is some concern that the pellets, which are 3-mm in diameter, may interfere to some degree with normal articulation, particularly at the tongue tip.

**Electromagnetic Midsagittal Articulometry (EMMA)**

Electromagnetic midsagittal articulometry (indicated \(EMMA\) in Table 2.1) records the movement of small electromagnetic receiver coils which are affixed to the articulators with dental adhesive (Perkell et al., 1992). The data thus derived are similar to x-ray microbeam data in their pointwise nature and relative ease of interpretation. With a 4X4-mm mounting surface and 2.5 mm height, EMMA transducer coils are also about the same size as x-ray microbeam pellets and subject to some of the same limitations. EMMA systems are less costly than x-ray microbeam systems but have additional sources of error if transducer coils are placed out of midline and happen to twist (i.e., rotate in the
transverse plane) during recording. With careful use, however, it is estimated that the spatial resolution of the system can be between 0.5 and 1.0 mm and the time resolution similar to that for x-ray microbeam at 100 Hz or higher.

*Ultrasound Imaging of the Tongue*

Although ultrasound imaging of the tongue has not been used previously to investigate the vocal tract inverse problem, it is described here because it is used in the current study. Ultrasound imaging of the tongue (Shawker, Sonies & Stone, 1984; Stone, 1990) is accomplished by placing an ultrasound transducer that sends and receives ultrasound waves underneath the chin of the speaker either in the transverse or sagittal plane. Sound waves emitted by the transducer reflect off of locations of sudden density change in the oral structures, are detected by the transducer, and converted to a video image. Because the most abrupt density change occurs at the tongue-air boundary, the most prominent reflection indicates the location of the tongue surface. Extracting the tongue contour lines is facilitated by the use of edge detection software, but still requires some experimenter intervention for each frame and is therefore somewhat tedious. Fully automatic edge detection methods are currently under development (Akgul, Kambhamettu & Stone, 1998).

Ultrasound thus captures tongue surface contours of a more continuous nature than EMMA or x-ray microbeam, but does not provide information about the movement of individual tongue points (i.e., given two contours it is not possible to determine precisely which points in one contour correspond to which points in the other). Although very posterior tongue surfaces can be recorded with ultrasound (as far back as the hyoid bone), the most anterior portion of the tongue tip (as much as about 1 cm) can be lost due to a shadow cast by the jaw or an air pocket beneath the tongue tip. Also, the tongue surface contour can fade or be partially lost as the angle of the tongue surface with respect
to the direction of ultrasound wave propagation becomes oblique and less reflective, or when the tongue body contacts the palate (in which case a contour representing the bony palate may appear). A principal component analysis (PCA)-based method of analyzing ultrasound images of the tongue that does not rely on the explicit extraction of surface contours is explored in the current study. This method is described in detail in the next chapter.

Another possible limitation of the ultrasound machine used in this study is the relatively slow temporal resolution of about 30 Hz. (The fastest machines have scan rates of 60 Hz.) This is not a serious limitation for many speech sounds because most muscle-induced speech movements have bandwidths below 15 Hz (Müller & McLeod, 1982), but aerodynamically influenced movements such as trills and plosives may not be captured precisely at this rate. Finally, ultrasound equipment is less expensive and more widely available than the previously described systems.

**Video Imaging of the Face**

The movement of the lips, jaw, and other externally visible structures can be captured simply with a video camera. The lips may be highlighted with makeup to facilitate the automatic extraction of lip edges (e.g., Lavagetto, 1995) or more sophisticated marker-tracking systems may be used (e.g., Vatikiotis-Bateson & Ostry, 1992). The current study does not extract lip boundaries or marker positions, but applies the same PCA-based method to lip images as it does to the ultrasound images of the tongue.

**Accuracy of Articulatory Estimations**

Column 3 of Table 2.1 indicates the accuracy of estimated vocal tract shapes in those cases where accuracy was verified with actual vocal tract measurements. There has
been little attention given to articulatory verification in studies using speech production models. Instead, listening to the speech synthesized from the articulatory estimate is considered to be a sort of indirect verification that the underlying shapes are reasonable. While this method of verification addresses the primary concerns of studies looking to use an inverse mapping mechanism in speech recognition and speech coding, it does not provide a reliable measure of articulatory shape estimations. This is because inconsistency in the inverse mapping (which can be more clearly demonstrated for production models than for humans) can cause the output of the synthesizer to match the input speech even though the articulatory configuration is quite different from that of the speaker. In cases where model output has been compared to human vocal tracts, the correspondence between actual and estimated shapes has been accurate only in a very gross sense. For example, Flanagan et al. (1980) report comparisons to area functions calculated from x-ray for the two distinct vowels /a/ and /i/. While the comparisons were reported as “gratifyingly close,” they contained area estimates off by as much as 6 cm$^2$ (or almost 200%) for some positions along the tract.

Some of the studies which learned a mapping based on human acoustic-articulatory data also evaluated the accuracy of the articulatory estimations indirectly by testing an end-product application that uses the estimates. Zlokarnik (1995) demonstrated that articulatory estimates could be incorporated into an automatic speech recognition system to reduce error rates from 10.6% to 8.7%. Lavagetto (1995) tested the “speechreadability” of animated lip images estimated from acoustics of single words and found that in some conditions as many as 64% of the words which were correctly identified from original movies could be identified from the estimation-based animations.$^2$ While these evaluation measures strike to the heart of what is important in

---

$^2$But this preliminary report is somewhat vague about the experimental details.
each study, they do not give a precise indication of how close the articulatory estimates were to the actual configurations or for which sounds errors were made.

Of the studies that have verified estimations against actual measurements, mean errors have tended to be in the range of a few millimeters, or about 4–14% of the range of motion of the studied articulator. Specifically, Ladefoged et al. (1978) report mean errors of 3.9 mm for the lips and tongue, which represents approximately 13% of the range of motion (assuming a range of variability of about 3 cm as estimated from Harshman et al., 1977, which covered the same data). Papçun et al. (1992) report root mean square (rms) errors for vertical pellet positions on the lips, tongue tip, and tongue dorsum) ranging from 8% to 14% of the range, and correlations of pellet trajectories ranging from .94 to .19, depending on the sound and the articulator being measured (with pellets on articulators that moved little having low correlations). Hogden et al. (1996) report rms errors of about 2 mm and correlations of .94 for four coils on the tongue (which was about 10% of the range of motion for the tongue-rear coil, the only coil for which the range was reported), with smaller errors (0.5–2 mm) but worse correlations (0.50–0.80) for coils on the jaw and lips. Root mean square error in the study by Zacks & Thomas (1994) was as low as about 4% of measured range for four pellets attached to the tongue and lips (though the speech corpus used in this study was arguably much easier than those used in the other studies). In general we should keep in mind that the speech materials and mapping methods used in these studies varied considerably; the numbers are presented not so much for comparison between studies, but rather to give some idea of the range of success that has been achieved.3

3Values of the percentage of the range of variation of the various articulators have been calculated from each study’s report of error and range of motion. Error values were not originally reported in this percent-of-range format.
INCORPORATION OF DYNAMICS

Column 4 of Table 2.1 indicates whether constraints from vocal tract dynamics were used to improve articulatory estimates, and if so, how. The reason for incorporating dynamics is simple: if articulatory trajectories are smooth and regular, we should be able to use knowledge of the smoothness and regularity to improve articulatory estimations. This type of extended temporal information is thought to be particularly important for handling stop consonants which contain a brief period of silence — a period at which estimation of articulation from acoustics alone must be impossible.

Most studies that have included dynamical constraints have used what can be called the minimal movement constraint, which says that the current shape should be similar to the previous shape, or to put it differently, that large sudden variations should be penalized. This simple constraint was incorporated in the analysis-by-synthesis model of Levinson & Schmidt (1983) by initializing the parameters of the articulatory synthesizer for the current speech frame with the final parameters from the previous frame. While this method did produce smooth motions when articulating diphthongs, it was not reported whether the motions would have been smooth without this special initialization — a possibility we must take seriously since it is also reported that “tests indicate that the method is not terribly sensitive to initial values of the articulatory parameters” (p. 1150). Whatever the effect on smoothness, the constraint seems to have actually increased rather than decreased articulatory estimation error by “freezing” certain articulators and forcing the model to try to produce good speech output by compensating with other articulators. The final result for some diphthongs was quite unnatural (e.g., almost no rounding of the /u/ in /ou/).

A different implementation of the minimal movement constraint simply smoothes the vocal tract estimates of the model post hoc (indicated smoothing in Table 2.1). This
procedure would seem to run the risk of averaging away from hard-won articulatory estimates, however. (Unfortunately, the accuracy of the estimates made using this technique was not verified.) Simple smoothing does in any case remove “glitches” in the output speech synthesized from the estimated vocal tract shapes (Rahim et al., 1991).

A more sophisticated implementation of the minimal movement constraint involves the use of dynamic programming (DP) to select a sequence of shapes that simultaneously minimizes both acoustic error and articulatory change (e.g., Schroeter & Sondhi, 1994). While improved smoothness of articulatory movement has been demonstrated with DP, there has been no analysis of whether shape predictions are actually improved by adding this constraint (smoothness does not guarantee accuracy), or indeed any articulation-based validation of the accuracy of the shape predictions at all.

While inclusion of the minimal movement constraint is a step in the right direction, we should be able to incorporate more knowledge of articulatory dynamics than simply that the current prediction should be similar to the previous one. The extended Kalman filter has been used in a few studies to attempt to incorporate a more detailed knowledge of dynamics (noted in Table 2.1 as Kalman). Kalman filtering methods are applicable when there is a physical system that changes over time and produces a measurable output which is a function of the state disturbed by noise. A Kalman filter is able to estimate the state of a system by alternating between estimates and measurements of the system, recursively conditioning the estimates on all of the past measurements.

While a standard Kalman filter works on linear systems, an extended Kalman filter works for nonlinear systems by iteratively approximating the system as linear in a small range. For the vocal tract inverse problem, the acoustic measurement is considered to be a noisy function of the underlying articulatory state. While Kalman filtering methods are attractive, they are sensitive to assumptions about the dynamics of the underlying system — and in the case of articulatory movements in the human vocal tract, these dynamics are
not well worked-out (Schroeter & Sondhi, 1994). They also depend on a speech production model for the estimates of the acoustics, and are thus limited by whatever inaccuracies may be present in that model. Unfortunately, studies which have used Kalman filtering methods have not provided articulatory validation.

A different way to incorporate temporal information is to use a wide acoustic window as the basis for the articulatory estimate (indicated wide window in Table 2.1).4 That is, in addition to using the current acoustic output as information about the current articulatory state, one can use the recent past and upcoming acoustic information as well. In this way, articulatory estimates are based on acoustic trajectories rather than on nearly instantaneous acoustic states. While this method does not directly make use of information about articulatory dynamics, it can be thought of as doing so indirectly to the extent that there is a regular relationship between acoustics and articulation.5 One drawback of using a wide acoustic window is the additional computational complexity it entails. For a neural network implementation of the mapping procedure this corresponds to having more input units (expensive but manageable), but for a lookup-table or statistical regression implementation the additional complexity could be prohibitive.

Another way to capture dynamics in a neural network implementation is to incorporate recurrent connections that feed information from a later stage of processing back to an earlier stage (Elman, 1990; Jordan, 1986; Rumelhart, Hinton & Williams, 1986). With such an architecture, the state of the network depends not only on the current

4The term wide window, as used here, should not be confused with using a long time frame in spectral analysis. For example, Papçun et al. (1992) used a wide window of about 200 ms composed of a series of (overlapping) 16-ms frames.

5While there is much debate about how regular the relationship is, that there is some regularity is not in question.
input but also on the previous state of the network. In this way temporal order is represented in the network implicitly by the effect that it has on processing. Recurrent neural networks have not been previously applied to the vocal tract inverse problem but are used in the current study.

**ACOUSTIC ANALYSIS METHODS**

Column 5 of Table 2.1 lists the methods used for analyzing the acoustic signal. All studies reviewed here have used spectral (frequency-domain) methods. Where actually measured, the spectral analysis is based on short windows (16-32 ms) of speech. (Studies that deal only with speech from models rather than with human speech need not perform spectral analysis because the spectral information is generated directly in the speech production model, e.g., Atal et al., 1978; Atal & Rioul, 1989; Boë et al., 1992.)

Many studies simply use the power spectrum itself (or its logarithm) as the acoustic representation (indicated $PS$ in the table). In some studies the power spectrum is represented on a Bark scale (indicated $Bark$), i.e., more precisely in the low-frequency range and less precisely in the high-frequency range in a manner similar to that found for the human ear (Zwicker & Feldtkeller, 1967). Other studies use formant frequencies and/or amplitudes and/or bandwidths. Studies which used a formant-based representation studied only static vowels or only synthetic speech, most likely due to the difficulty of accurately and reliably estimating formant information for normal continuous speech.

Several studies have used representations based on linear predictive coding ($LPC$). Predictive methods of speech coding can be efficient because, given a reasonable prediction, the dynamic range of the error (the difference between the prediction and the actual signal) can be much smaller than that of the original signal, and therefore require fewer bits for accurate encoding. In LPC, the speech sample at a given time is approximated as a linear combination of some number (usually about 8–12) of the
previous speech samples. The only information that needs to be transmitted is the set of coefficients that describes how to combine each of the previous samples for a new estimate, plus the error that this estimate produces (which, again, should have a small dynamic range). Of course if a new set of coefficients had to be sent with each sample, then there would be no savings (but great cost). But the coefficients that describe how to combine the previous samples are related to the shape of the vocal tract (specifically, to its transfer function) and the nature of the source (either periodic, as in voiced speech, or random, as in unvoiced speech) and thus change relatively slowly. They can therefore be updated less often (typically about every 10 ms, whereas the signal itself may be sampled about once every 0.1 ms or faster). Importantly for speech analysis, the predictor coefficients may be estimated directly from the speech signal so that they describe a digital filter with spectral properties similar to those of the original speech waveform (Markel & Gray, 1976; Rabiner & Juang, 1993; Tempelaars, 1996).

An important feature of LPC analysis is that it separates out the effects of the source (the glottal impulse or noise), and the filter (the vocal tract). In fact, if the presumed model were accurate, such that the effects of the source were completely independent of the effects of the vocal tract, and if perfect predictor coefficients could be found, then the source would be perfectly described by prediction error and the transfer function of the vocal tract would be perfectly described by the predictor coefficients. While neither of these assumptions hold in practice, it turns out that the predictor coefficients produce a spectrum that is similar to, but smoother than, the log power spectrum (as normally derived by Fourier analysis). As indicated in the table, some studies have used this LPC spectrum directly, while others have used a numerically smoothed power spectrum (Hogden et al., 1996). Other studies have use the LPC coefficients themselves as an efficient representation of the acoustics.
Still other studies have used *cepstral coefficients*, which are the coefficients of the Fourier transform representation of the log power spectrum. These may either be derived from the LPC coefficients (Rabiner & Juang, 1993) or calculated from the speech signal using Fourier transforms (by taking the inverse Fourier transform of the log of the Fourier transform of the signal). Cepstral analysis is related to LPC analysis in that it also separates the effects of the source and vocal tracts. It can be understood simply as treating the log power spectrum, which is really a representation in the frequency domain, as a signal in the time domain. Viewed in this light, the log power spectrum can be said to have a high frequency\(^6\) component (related to the source) imposed on a lower frequency component (related to the vocal tract). By performing Fourier analysis as if this were a time domain signal, the high and low frequency components (source and vocal tract) can be separated.

It should be pointed out that both LPC and cepstral analysis are most accurate for the quasi steady state voiced regions of speech and less so during the unvoiced and transient regions (Rabiner & Juang, 1993). More than one study has started out using cepstral or LPC coefficients and found that a simple power spectrum representation worked better (Strube and colleagues, as reported in Schroeter & Sondhi, 1994; Hogden et al., 1996). This was also found to be the case in the current study.

**RANGE OF SOUNDS STUDIED**

The final column of Table 2.1 indicates which speech sounds were studied in each experiment. Notice that most studies have examined only vowels. The reasons for this are several. First, vowels are the easiest sounds to produce with a speech production

---

\(^6\)Because the log power spectrum is not really a signal in the time domain, this has been called “quefrency” rather than “frequency.” Note that the word “cepstral” is itself a play on the word “spectral.”
model, an essential element in most studies. Particularly if the model has only a single cavity, nasal-like sound production is difficult, and estimating a shape similar to the human vocal tract for such sounds is clearly impossible. As indicated in Table 2.1, some studies did include a nasal parameter in their model and thus were able to examine nasals.

A second reason studies have often avoided consonants is that many consonants demand some incorporation of dynamic constraints, which most models have not included. The brief silent portion of a stop gives no information for estimating articulation from static (short-window) acoustics, for example. Transitional cues — changes over time before and after the stop — are thought to be the primary source for identifying place of articulation for stops.

A third reason for the emphasis on vowels involves the method of acoustic representation. A representation in terms of formant frequencies is not appropriate for most consonants as the formants may be ill-defined (consider fricatives, for example, with a broad band of energy in the high frequency range). Also, LPC-based methods may be more appropriate for voiced than unvoiced sounds, as was pointed out above. Finally, cepstral analysis was devised as a method for separating out the contribution of source and filter, a notion that makes less sense when the source is broad-band noise (as in fricatives). When considering the choice of acoustic representation, it seems that most studies have been focused on vocalic sounds from the outset.

The only speech production model-based study to examine consonants was that by Schroeter & Sondhi (1994), and this system initially failed on stops due to the silence and burst. Several ad hoc modifications specifically designed for stop consonants were used to improve this performance. For example, two different dynamic programs were run: one up to the moment of silence and one from the onset of voicing after the burst. Even after such improvements, on a test using 204 spelled letters recorded by 4 speakers the point of closure was estimated incorrectly 35% of the time (as measured by “visual
inspection,” not by comparison to direct measurements). There were no “obvious” errors on the vowels.

Several of the studies using measured human vocal tract data have included consonants in the speech materials. Zacks & Thomas (1994) and Hogden et al. (1996) used consonants only as a frame for the vowels, however, and did not include any contrastive consonants that the system would have to learn to distinguish. Zlokarnik (1995) included as many as 15 consonants and 4 vowels in a $V_1CV_2$ format, but did not distinguish (at least in the published abstract) how well the system performed on consonants versus vowels. Lavagetto (1995, 1997) included some consonants in some portions of his study on lip estimation but also did not report separate consonant and vowel performance rates, though he did indicate “particularly severe” errors on stops. The study by Papçun et al. (1992) on stop consonants is thus the only study to directly address the issue of estimating articulatory positions during consonant production. The wide acoustic window used in this study captured transient information on either side of the stops and estimated articulatory configurations quite accurately.

**Background Summary: Where more research is needed**

Through this overview of techniques that have been applied to the vocal tract inverse problem we can obtain a clearer view of where more research is needed. Because the simplifying assumptions in speech production models have lead to uncertain and unverified results, further study of real human vocal tracts is called for (in part so that more realistic production models can be developed). Although it is difficult to measure human vocal tracts in real time without disrupting normal speech, several acceptable methods of measurement have been developed, including x-ray microbeam, electromagnetic midsagittal articulometry, and ultrasound and video imaging. Other methods may become available in the near future as well; for example magnetic
resonance imaging methods are promising (Alwan, Narayanan & Haker, 1997; Stone, 1991), but still too slow to capture important speech movements. The current study uses ultrasound and video imaging and explores an analysis method based on the principal component analysis of raw images which avoids the need to perform explicit edge extraction.

In recent years more attention has been paid to incorporating dynamic aspects of articulatory motion into solutions of the vocal tract inverse problem. While the incorporation of a “minimal movement constraint” produces smoother articulatory output, use of more detailed information of typical acoustic-articulatory trajectories offers greater promise for enhancing estimation accuracy. Such trajectories may be captured to some degree through the use of wide acoustic windows (i.e., by simultaneously analyzing several acoustic analysis frames), and potentially by extended Kalman filters and recurrent neural networks. These techniques have been applied in only a few instances to the vocal tract inverse problem, however, and their potential contribution to this area remains largely unknown. In fact, the current study represents the first application of recurrent neural networks to the vocal tract inverse problem.

Finally, study of the vocal tract inverse problem has been overwhelmingly centered on the study of vowels as opposed to consonants. Many of the inverse problem methods that have been developed apply to consonants either not at all or with considerably reduced accuracy. Studies that have included consonants in the speech materials have often not analyzed the distribution of errors across sound categories. Stop consonants are particularly difficult from an inverse problem point of view because they may contain periods of very low acoustic energy and very rapid articulatory movements. In this sense they demand attention to dynamics. For this reason, the current study analyzes a small corpus of stop consonants and a larger corpus composed of nine consonants and eight vowels.
CHAPTER 3: LEARNING THE DYNAMICS OF VOICED STOPS IN A SCHWA CONTEXT

This chapter presents the initial attempt to map speech acoustics to video and ultrasound images of the lips and tongue using a recurrent neural network. It is divided into two parts. Part one describes the articulatory analysis method: the use of ultrasound, video, and principal component analysis (PCA) to capture and efficiently describe tongue and lip movements. It is shown that these analysis techniques can be used not only to investigate the vocal tract inverse problem, but also to directly study a wide range of speech production phenomena such as coarticulation, articulatory variability, and production asymmetries. Part two describes the use of the PCA-based representation to train neural networks to estimate tongue and lip shapes from speech sound patterns.

Part One: Principal Component Analysis of Lip and Tongue Images

METHODS

Data Collection

The speech corpus consisted of the three utterances /bɒbɒbɒbɒb/, /dɒdɒdɒdɒd/, /gɔgɔgɔgɔ/. The entire corpus was spoken three times for a total of nine utterances. All utterances were spoken by a male native English speaker (the author). Vocal tract movement and acoustic data were simultaneously collected during each utterance.

Images of the tongue were recorded with a 2-4 MHz variable frequency ultrasound scanner (Performa model; Acoustic Imaging, Phoenix, AZ). The transducer was placed in the sagittal plane underneath the chin of the speaker. Head movement was restricted and the transducer was held in position using a specially designed head and transducer support system (HATS; Stone & Davis, 1995), as shown in Figure 3.1. A
gelatinous, acoustically transparent standoff of polymerized mineral oil (Kitecko; 3M, St. Paul, MN) was placed between the transducer and the jaw in order to allow freedom of jaw movement. The standoff, which is more compressible than the soft tissue of the chin, has been found to reduce soft tissue deformation to undetectable levels (Stone & Davis, 1995). Slice width was 2 mm at focal depth. The ultrasound scan rate was 28 scans per second, with two frames repeated each second for a standard video output of 30 frames per second. Estimated measurement error of this recording method is 0.5 – 0.7 mm (Stone, Shawker, Talbot & Rich, 1988). The video output of the ultrasound scanner was sent to a digital AV mixer (model WJ-MX30; Panasonic, Secaucus, NJ) for combination with the other recorded signals.

Figure 3.1: The HATS system for ultrasound recording of the vocal tract (Stone & Davis, 1995). (Facial makeup was not used the first study.)

A frontal video image of the face was recorded with a digital Hi-8 video camera (model VM-H720A; Hitachi, Nocross, GA) placed on a stable mount. Side lighting was used so that shadows would help define facial features. The 30-frames-per-second video output was inserted into a portion of the ultrasound image using the video mixer.

The subject’s speech was recorded with a unidirectional short-range microphone (model AT857AMa; Audio-Technica, Stow, Ohio) mounted on the HATS system and held a few inches from the mouth of the subject. The recording environment was subject to various sources of noise, including the fan on the ultrasound scanner (attenuated to
some degree by acoustic insulating foam placed between the scanner and the subject), occasional noise from adjacent rooms, and ambient room noise such as that from the air conditioning system, etc. The analog audio signal was inserted into the hi-fi audio channel of the VCR.

**Data Processing**

The audio-visual signal on the video tape was captured and digitized on an Indigo 2 XL computer equipped with a Cosmo Compress video compression board (Silicon Graphics, Inc., Mountain View, CA). Video was captured at a rate of 29.97 frames/sec and audio at a rate of 22,050 Hz. To reduce onset and offset effects, the middle three syllables from each five-syllable utterance were extracted and used. Utterance boundaries were determined by reference to a visual plot of the audio signal and confirmed by listening to the visually selected portion. Using these boundaries, digital movies of individual utterances were extracted and saved.1 The captured corpus contained a total of 202 video frames which were converted to grayscale images. The audio and video signals were then processed independently to put them into a format suitable for presentation to a neural network.

---

1The images in this corpus were captured with video field interlacing turned off. When the digitized movies were then converted to a series of still frames, fine striations appeared in some images (most noticeable when the image was part of a fast movement sequence) even though the interlacing setting used in the movie conversion software was also set to “none.” For subsequent corpora, interlacing was set to “odd fields first” during movie capture and conversion to stills and the striations disappeared.
**Audio Processing**

The audio signal was subsampled at 11,025 Hz to reduce the computational load. The signal was *pre-emphasized* in order to boost the energies at high frequencies and thereby produce a flatter signal spectrum. The output, $S(n)$, of the pre-emphasis filter at time $n$ was defined by

$$S(n) = s(n) - 0.9375s(n-1)$$

where $s(n)$ was the original subsampled signal (after Rabiner & Juang, 1993). The pre-emphasized signal was then blocked into 50% overlapping frames of 22.2 ms (245 samples). This resulted in a 11.1-ms audio frame rate which could be synchronized to the 33.3-ms video frame rate at a 3:1 ratio. In order to minimize signal discontinuities at the edges of frames, a Hamming window was applied to the signal in each frame. The Hamming window had the form

$$w(n) = 0.054 - 0.46 \cos \left( \frac{2\pi n}{N-1} \right), \quad 0 \leq n \leq N-1$$

where $N$ was equal to the number of samples in the frame. The result of windowing each frame was the signal

$$x(n) = S(n)w(n), \quad 0 \leq n \leq N-1.$$  

The log-LPC spectrum was computed for each frame using the autocorrelation method (Rabiner & Juang, 1993; Markel & Gray, 1976).

The log-LPC spectrum for each audio frame was partitioned into 18 bins, whose borders were determined as follows. First a bark scale was approximated by defining $\beta$ to be a function of frequency, $f$ (in Hz), such that

$$\beta(f) = 1, \text{ for } f < 1000$$

$$= \frac{(5500 - f)^2}{4500^2}, \text{ otherwise.}$$

---

2For $n = 1$, $s(n-1)$ was set to $s(n)$. 
See Figure 3.2 (a). When the area under this curve in the region from 250 to 5500 Hz\(^3\) is divided into 18 bins of equal area, the bin borders roughly approximate the locations of critical bands found in human listeners (Rabiner & Juang, 1993; Zwicker & Feldtkeller, 1967). Instead of using these bins directly, however, \(\beta(f)\) was filtered by \(l(f)\), the detrended mean log-LPC spectrum of the entire corpus (normalized to range from zero to one) to form

\[
B(f) = \beta(f)l(f)
\]

which was then divided into 18 bins of equal area, as illustrated in Figure 3.2 (b). The effect of this manipulation was to concentrate the bins near the information-bearing peaks in the acoustic data and thus be able to represent the most important aspects of the acoustics more precisely with a small number of bins (which was limited by the need to keep the number of input units to the network small) but still with some sensitivity to the bark scale.\(^4\) The bin borders produced by this method are compared to the standard bark scale in Figure 3.2 (c) and (d). The mean of the log-LPC spectral magnitude was computed for each bin. These values, which will be referred to as spectral bins, were normalized over the entire corpus to range from zero to one and served as the input values for the networks.

As a summary, Figure 3.3 provides a graphical display of the several steps of the audio processing procedure on the utterance /gəgəgə/.

\(^3\)Little distinguishing information was expected outside this frequency range. See Papçun et al. (1992), Rabiner & Juang (1993).
Figure 3.2: Spectrally-weighted bark bins: (a) the function $\beta(f)$ is used to approximate a bark scale (see text); (b) $\beta(f)$ (dotted) is filtered by the mean log-LPC spectrum of the entire corpus (dashed) to create $\hat{B}(f)$ (solid), the area under which is divided into equal parts; (c) an expanded view of the spectrally-weighted bark bins from (b); (d) standard bark bins from human experiments shown for comparison (Zwicker et al., 1957).

**Video Processing**

Position and Intensity Normalization

Position and intensity normalization for the lip images were carried out with respect to a relatively stationary *reference area* of the face image containing the bridge of the nose and clear lines defined by the subject’s glasses. One image was arbitrarily chosen as the *reference image* and all other images were adjusted so that their reference area achieved a closest match with the reference area of the reference image.

\footnote{Note that with a more complex speech corpus, the mean log-LPC spectrum would be flatter, with the energy distributed more evenly across all speech-typical frequencies. With a perfectly flat mean spectral content, this procedure would revert back to the simple bark scale approximation.}
Position normalization: Each image was normalized by shifting the reference area of the image either not at all or a single pixel up, down, left, or right and in each case comparing the goodness of fit to the reference area of the reference image. Goodness of fit was determined by finding the two dimensional correlation coefficient, \( r \), between the matrices defined by the two reference areas. If \( A \) is the matrix of intensity values representing the reference area of the reference image and \( B \) is the matrix of intensity values representing the (possibly shifted) reference area of image \( i \), then \( r \) was defined as:

\[
r = \frac{\sum \sum A_{jk} B_{jk}}{\sqrt{\sum \sum A_{jk}^2 \sum \sum B_{jk}^2}}
\]

where \( j \) and \( k \) are the row and column indices of each matrix. If the highest correlation was found with no shift, then no movement relative to the reference image was inferred.
If the highest correlation was found with some shift, then the image was considered to have moved relative to the reference image (due to subject movement, camera movement, or both), and the search was repeated, taking the shifted reference area as the new reference area, until no more movement was found. Across the entire corpus, no horizontal movement and vertical movement of only a single pixel (about 0.8 mm) was detected. When the lip images were adjusted to compensate for the detected movement the movies seemed to jump up and down slightly. Therefore, the slight detected vertical movement was presumed be due to movement of the subject’s glasses across a pixel boundary rather than global movement of the head and no correction for movement was eventually made. The minimal detected movement indicates that the HATS system was very stable.

**Intensity normalization:** The mean intensity of the reference area was found to change abruptly by 5–10% a few times during the recording session. These changes were assumed to be caused by the auto-iris in the video camera. These sudden intensity shifts were corrected by scaling the pixel intensities of each original_image, by the ratio of $INT_{ref}$, the mean intensity of reference area of the reference image, to $INT_i$, the mean intensity of the reference area of image:

$$normalized\_image_i = original\_image_i \left(\frac{INT_{ref}}{INT_i}\right).$$

Cropping and Scaling

After normalizing, rectangles containing the lip and tongue images were cropped from each video frame. A 2:1 aspect ratio was used in each case, with the tongue images 200 X 100 pixels and the lip images 100 X 50 pixels. The images were then scaled to 80 X 40 pixels, a size small enough to be computationally manageable in subsequent processing, but large enough to view details easily.
Principal Component Analysis

PCA was performed separately on the sets of lip and tongue images. Each cropped image was converted to an image-vector, $\Gamma_i$, by reading pixel intensities row-wise down the image. The average-image-vector

$$\Psi = \frac{1}{n} \sum_{i=1}^{n} \Gamma_i$$

was then subtracted from each image vector to form the set of average-subtracted-image-vectors

$$\Phi_i = \Gamma_i - \Psi.$$  

Defining the matrix $A = [\Phi_1 \, \Phi_2 \, ... \, \Phi_n]$, the principal components of the set of images would normally be found by obtaining the eigenvectors, $u_k$, and corresponding eigenvalues, $\lambda_k$, of the covariance matrix

$$C = \frac{1}{n-1} \sum_{i=1}^{n} \Phi_i \Phi_i^T = \frac{AA^T}{n-1},$$

which satisfy the equation

$$\frac{AA^T}{n-1} u_k = \lambda_k u_k.$$

Because $AA^T$ is very large, however, the computationally more tractable method suggested by Turk & Pentland (1991) was used in which the eigenvectors, $v_l$, of the much smaller matrix

$$\frac{A^T A}{n-1}$$

were found and sorted according to their eigenvalues. The first $n$ desired eigenvectors, $[u_1 \, u_2 \, ... \, u_n]$, were then calculated from these “intermediate” eigenvectors, $V = [v_1 \, v_2 \, ...$ 

\[5\] The size of $AA^T$ is the square of the number of pixels per image ($3200^2$), while the size of $A^T A$ is just the square of the number of images ($202^2$).
\[ \begin{bmatrix} u_1 & u_2 & \ldots & u_n \end{bmatrix} = AV. \]

See Turk & Pentland (1991) for the justification of this technique.

The eigenvectors, or principal components (PCs), \([u_1 u_2 \ldots u_n]\) were then reshaped into 80 X 40 matrices and rendered as grayscale intensity values to form the eigenimages of the original image set. The principal component loads, \([\omega_1 \omega_2 \ldots \omega_n]\), for each original image, \(\Gamma\), were then calculated by
\[ \omega_i = u_i^T (\Gamma - \Psi) \]
for \(i = 1,\ldots,n'\), where \(n'\) is the number of principal components retained. Typically, \(n' << n\) because the higher order principal components (those associated with higher eigenvalues) contain most of the differentiating information (i.e., explain most of the variance in the dataset), and the lower order components can be discarded without much information loss. While the relative magnitude of the associated eigenvalues indicates the relative amount of variance explained by each component, it was not clear in this study how many PCs would be needed to be able to reconstruct perceptually convincing lip and tongue images without actually reconstructing several images with various numbers of components. An image was reconstructed from \(n'\) components by
\[ \Gamma_{\text{recon}} = \Psi + \sum_{i=1}^{n'} \omega_i u_i. \]
Several lip images from a /bəbə/ utterance and several tongue images from a /dədə/ utterance were reconstructed with \(n' = 1, 2, 6, 10, \) and 20 and compared to the original images in order to determine \(n'\).

Because the loads were to be used as target values for a neural network with log-sigmoid output units, they were normalized over the entire corpus to range from 0.1 to 0.9. The pre-normalized range was saved to allow for reconstruction of images from network outputs. Finally, in order to match the video frame rate (30 fps) to the faster
audio frame rate (90 fps), two loads were non-linearly interpolated between each pair of original loads using a lowpass interpolating filter.

RESULTS AND DISCUSSION

Principal component analysis (PCA) of the ultrasound and video images produced compact representations of the lips and tongue not only suitable for training a neural network on the vocal tract inverse problem but also useful for the direct analysis of several speech production phenomena, including coarticulation, articulatory variability, and speech production asymmetries. The nature of the articulatory descriptions afforded by the eigenimages differs considerably from that allowed by more traditional representations in which points on the surface of various articulators are tracked. Instead of reporting results in terms of millimeters of movement of an articulator, results are reported in terms of changes in the degree to which eigenimages describe an articulatory state. Because this type of representation may be unfamiliar, an in-depth description of the extracted principal components is presented first, followed by an analysis of the subtle variations in articulation they are able to track.

The Images and Their Principal Components

Sample images covering the point of maximal constriction to the point of maximal opening in the three consonants are shown in Figure 3.4. Note that in /bʊ/ the lips move from closed to open with little movement of the tongue, and in /da/ and /ga/ there is a lowering of tongue tip and tongue body, respectively, with little movement of the lips. The average lip and average tongue images for the entire corpus are shown in Figure 3.5.

The first seven PCs of the lip and tongue images are shown in Figure 3.6. In this figure each PC is displayed with maximal contrast — that is, with the maximum and minimum vector element values within each PC assigned maximum (white) and minimum
Figure 3.4: Sample images from the production of /ba/, /da/, and /ga/. Consecutive frames are shown from complete consonantal closure to the point of maximal opening.

Figure 3.5: The average lip and average tongue images (tip to the right).

While this method of displaying the PCs highlights what parts of the image each PC is tracking, it makes variations found in the lower order PCs seem more important than they actually are. The lesser importance of the lower order PCs is indicated by the fact that the loadings they assume for the reconstruction of the original images tend to be relatively small.

Figure 3.7 corrects this misrepresentation by showing the principal components scaled by the range of loading values they assume and normalized for brightness as a group instead of individually. While this representation shows the relative importance of the variation captured by the PCs appropriately, it tends to obscure the particular variation captured in the lower order components.
Determining the Number of Principal Components for Use in Network Training

Often in PCA, the number of PCs to retain for analysis is determined by the amount of variance explained by each component. This measure is obtained by dividing the eigenvalue of each PC by the sum of all of the eigenvalues. Figure 3.8 shows the percent variance explained by the first 20 PCs. For the lip images, over 80% of the variance is explained by the first three PCs; but for the tongue images, less than 50% of the variance is explained by the first three PCs, and even 20 PCs explain only about two-thirds of the variance.
It was not clear from this variance measure alone, however, how many PCs would be needed to produce images of good perceptual quality. This was particularly true for the tongue images because there was no interest in reconstructing the “snow,” from which a large portion of the variance was expected to arise. In order to more directly ascertain how many images would be needed for quality reconstructions, several images of the lips from a /bɔbɔbɔ/ utterance and several images of the tongue from a /dɔdɔdɔ/ and /gɔgɔgɔ/ utterance were reconstructed from 1, 2, 6, 10, and 20 PCs and the reconstructions were visually compared to the original images in both still and movie format. In addition, a mean pixel error score was computed between the original and reconstructed images across all images in the corpus.

The results of this analysis for the lips and tongue are shown in Figure 3.9 and Figure 3.10, respectively. Lip images that closely mimic the opening and closing of the mouth in /bɔbɔbɔ/ were reconstructed from only two PCs. For the tongue, the images reconstructed from only a few PCs differed from the original images in that the “snow” was not present and the tongue surface contour was not as crisp. The lack of a crisp
Figure 3.9: Reconstructions of lip images from various numbers of PCs. Each image set consists of eight consecutive frames from peak to peak closure in a /bobobob/ utterance. Original images are shown on the left, reconstructed images in the center, and difference images on the right. The graph at lower right shows how mean pixel intensity error (calculated over the entire corpus) drops as a function of the number of PCs used in reconstruction.

contour, especially at the tongue tip, persisted for some images reconstructed from as many as 20 PCs (note for example the third frame in the 20-PC set in Figure 3.10). But the basic front raising and back raising motions of the tongue which allow a discrimination of the /bo/ and /do/ and /go/ utterances were clearly apparent, especially in movie format, when images were reconstructed from only two tongue PCs. Following this analysis, it was decided to start small and use only two lip and two tongue PCs for initial network training.
Articulatory Features Represented by the Principal Components

In order to clarify the role that each PC plays in articulatory representation, Figure 3.11 and Figure 3.12 show images reconstructed with principal component loads at their maximum and minimum values. (Only the indicated PC is used for reconstruction in each case, so that each image amounts to the average image plus the weighted eigenimage.) It is clear from Figure 3.11 that the first lip principal component, lip PC1, captures the major motion of opening and closing the mouth. Lip PC2 captures variation in the images related to the teeth, which are made more visible when the upper lip tightens and raises. Lip PC3, in addition to tracking a difference in teeth visibility, tracks
an asymmetry in the opening of the mouth: when weighted by the minimum load, the lower lip appears lower on the speaker's left than the right, whereas the lips appear more nearly symmetric when this PC is maximally loaded. This asymmetry is also indicated above in Figure 3.6 and Figure 3.7 which show darkness on the left and brightness on the right. The lower order PCs capture increasingly more imperceptible differences.

Figure 3.11: Lip images reconstructed by setting the loading on each principal component to its maximum (top) and minimum (bottom) values.

In Figure 3.12 it can be seen that Tongue PC1 tracks back raising, with the maximally loaded image representing the /g/ tongue position (recall from the sample images in Figure 3.4 that the tongue surface is dim when the back is raised) and the minimally loaded image representing the relaxed schwa tongue position. Tongue PC2 tracks front raising, with the tongue tip maximally raised when this PC is minimally loaded. Notice that the back of the tongue moves forward in the throat as the tongue tip rises (cf. Figure 3.4) and that this simultaneous activity is also captured in PC2.

Figure 3.12: Tongue images reconstructed by setting the loading on each principal component to its maximum (top) and minimum (bottom) values.
Tongue PC3 is somewhat more complex than the PCs described above. This PC tracks most strongly the variation in tongue position over the course of the syllable /gə/, showing the difference between extreme back-raising and moderate back-raising (as found in the coarticulated schwa of /gə/). The lower bright white line in this pair of images results from reconstructing a /gə/-like image without the use of Tongue PC1, which would normally be highly loaded throughout the /gə/ syllable. This “extra” line serves to illustrate the point that each of the PCs cannot always be expected to track a single recognizable movement by itself. As will be illustrated below, however, an inability to precisely interpret the contribution of individual PCs does not limit the ability to use groups of PCs together to track motion, discover variability, find asymmetries, and even estimate shapes from acoustics.

The remaining tongue PCs capture other sorts of variability. PC4 captures a slight change in tongue body height. PC5 and PC6 capture fronting of the rear of the tongue. As with the lips, the lower order PCs capture ever more subtle differences.

**Articulatory Motion Represented by the Time-Varying Principal Component Loads**

Once the PCs have been found, speech production can be tracked by looking at how the loads on each PC change over the time course of an utterance. The relationship between articulation and PC loads is more complex, however, than that obtained by direct measurement of physical movement. This is because a change in one PC load may represent a variety of physical changes over the entire articulatory structure. Also, not all PCs are of equal importance, so that a large change in one load may represent less movement than a small change in another. Nevertheless, variation in PC loads does correspond in interpretable ways to variation in articulation. Several examples will be
provided here to illustrate this point. The examples show further that analysis of these loads can provide insight into articulation that might otherwise be missed.

Figure 3.13 shows how the loads on the first four lip PCs change over time for three sample utterances. The most obvious characteristic of this plot is that there is much more lip activity in /boʊboʊ/ than in /dədədə/ or /ɡəɡəɡə/, as expected. The plot also brings to light several more subtle and perhaps unexpected phenomena that might be missed in direct observation of the utterances. First, although the lips do not move much in /dədədə/, they move more than in /ɡəɡəɡə/. Second, although the three repetitions of the syllables are mostly consistent, there is some variability. This is seen most clearly in that the load on PC3 moves down rather than up in the third /boʊ/ syllable. Careful visual inspection of the images showed that there was indeed a slight difference in production in which the lips open slightly more in the third /boʊ/, revealing more of the teeth and more of the apparent asymmetry in the lips (Figure 3.14).

Figure 3.13: Variation in lip PC loads over time for three sample utterances.
Figure 3.14: Variability in the production of /bə/. Frames are taken from the second and third /bə/ syllables of the /bəbəbə/ utterance in Figure 3.13. The lips open slightly more in the third /bə/.

Another detail of speech production that Figure 3.13 helps us find is a temporal asymmetry in movement as the lips open and close. One might expect the motion of the lips to be the same in closure as in opening, only with the movement direction reversed. Such would be the case if the face were a rigid structure with opening caused by a single hinge at the jaw. But the asymmetrical variation on the load of lip PC4 indicates that this is not the case. The load on PC4 is at a peak during the opening of the lips, and is much reduced as the lips move toward closure. The physical nature of the asymmetry can be seen in Figure 3.15. In this utterance, lip movement initiates near the midline, with motion of the lateral labial structures lagging slightly behind. On opening, the corners of the mouth tend to stay closed until the midline structures are far apart; on closing, the corners of the mouth are already open and only close at about the same time as the midline. The difference in opening and closing is highlighted by comparing images of the lip opening gesture to images of the lip closing gesture with the sequence reversed. Although this difference is subtle, the asymmetric behavior of the load on lip PC4 makes
it quite obvious. (Although a faster frame rate would be desirable, this effect was verified in subsequent experiments, as discussed below.)

Let us turn now to tongue motion. The behavior of the tongue PC loads is shown in Figure 3.16. In this figure it is clear that, as expected, most tongue motion occurs in the utterances /dədə/ and /gəgə/. A slight raising and lowering of the tongue body in /bɔbɔ/ is captured in PC4, however. This motion is probably caused by the opening and closing of the jaw, which serves as a support structure for the tongue. Slight variability can also be seen across the three syllables of each type. For example, PC1 drops lower (indicating more lowering of the tongue rear) in the first /gə/ than in the final /gə/.

Another aspect of tongue motion becomes apparent when the loads are plotted in the two dimensional space defined by PC1 vs. PC2 (Figure 3.17). In this space, the three utterance types are well separated even though they all contain the same vowel /ə/. This indicates that the schwa is highly coarticulated with the neighboring consonant. In the figure, the tongue images corresponding to some points are shown in order to give an articulatory interpretation to the various locations. It is notable that the rear of the tongue...
remains relatively forward throughout the /dødødə/ utterances and that the tongue body remains relatively high throughout /ɡɡɡɡɡə/.

Figure 3.16: Variation in tongue PC loads over time for three sample utterances.

Figure 3.17 also shows a small amount of variability within each utterance type, especially for /dødødə/ and /bɔbɔbə/. For /dødødə/ there is a notable difference among the repetitions in the shape of the tongue rear upon tip raising. The small differences in the three /bɔbɔbə/ repetitions seem to represent only very slight differences in tongue position, however. It is important to keep in mind that distances in this space do not correspond linearly to distances in physical space. Of particular importance for this figure is the fact that schwa-like configurations (the upper left region of the figure) are probably more precisely represented in this space because a higher proportion of the images in the corpus contained schwa-like shapes than shapes of back and tip raising (because all of the utterances contained schwas, but only some velar or alveolar closures). The difference in
Figure 3.17: Tongue load trajectories in PC1-PC2 space for all nine utterances. The three utterance types (/bɔbɔba/, /dɔdɔda/, and /ɡɔɡɔɡa/) are distinctly separated in spite of the common vowel, indicating coarticulation. The three individual utterances of each type are distinguished by line type (first repetition: solid; second repetition: dotted; third repetition: dashed). Stars indicate points for which an image is shown.

representation precision come from the fact that PCA is sensitive to overall variability, which can be caused by a few large changes or many small ones. The small variations in the large number of schwa images may therefore be more precisely represented than the large variations in the smaller number of raised-tongue images. This is one reason that
findings made on the basis of analysis of principal component loads should be verified by direct inspection of the images underlying the analysis. This notion is discussed in more detail below.

When the analogous plot in PC1-PC2 space is made of lip movement, no clear separation of utterance types occurs; because there is excessive overlap the analogous
plot is not shown for the lips. Instead, each utterance type is plotted in a separate quadrant in Figure 3.18, with the articulatory interpretation of points in the space given in the lower right quadrant. This plot illustrates variability across utterances of the same type, especially in /dɔdəð/ and /ɡəɡəɡə/, where the lips are more widely opened in two of the three utterances.

A caveat is in order with respect to the variability illustrated in Figure 3.18. Because the teeth are white in an otherwise dark oral cavity, opening the mouth wide enough to show the teeth introduces a nonlinearity in the relationship between pixel intensity values (to which PCA is sensitive) and relative lip aperture (what we really want to measure). That is, at a certain point of aperture, a slight opening causes a large sudden change in pixel intensity values (as the teeth appear). This may cause the variability in lip aperture to be somewhat exaggerated. (This is one of the reasons why the teeth were blackened in the following experiments.)

**Summary of PCA Results**

PCA applied directly to vocal tract images differs from traditional articulatory representational techniques in that vocal tract shapes are described in terms of PC loads instead of pointwise positions. The PCs tend to capture global aspects of movement of the articulators. In this corpus of voiced stops in a schwa context, the first two lip PCs tracked lip aperture and the first two tongue PCs captured aspects of back raising and front raising, respectively. The lower order PCs were found to track more subtle but systematic aspects of movement, such as asymmetries of lip aperture (lip PC3), midline vs. lateral lip aperture (lip PC4), and, for the tongue, intermediate levels of back raising (tongue PC3), and slight tongue movements in /bɔbɔbɔ/ utterance caused by the underlying movement of the jaw (tongue PC4). Variations in the loads on these PCs were found to correspond to observable articulatory variations and were found to
highlight certain aspects of speech production such as temporal asymmetry of motion, production variability, and coarticulation.

Although eigenimage analysis was initially used simply to find a compact representation of the lip and tongue images suitable for training a neural network, the method may be useful as a tool for speech production analysis generally. Because the method can be fully automated, it makes practical the analysis of large amounts of articulatory data produced by video and ultrasound recording, two of the most convenient and least invasive vocal tract measuring methods available. One limitation of the method is that distances in PC space do not correspond directly to distances in physical space, which may cause the variations in the loads to exaggerate or underestimate real variations in articulatory motion. Although exaggeration of small but systematic variations may sometimes be beneficial in that it magnifies variations that might otherwise go unnoticed (such as the temporal asymmetry in lip aperture in this study), for many studies of articulation this would be considered an undesirable trait. It is suggested, however, that the PC-based representation could be converted to a standard Cartesian representation by manually measuring articulatory surface contours of images at key points in PC space and numerically mapping the PC loads of the remaining images to the Cartesian space by nonlinear interpolation. Future studies are planned to assess the accuracy of such a method. This method might also be combined with automated edge detection algorithms, which to date have had only limited success with ultrasound images of the tongue. Depending on the goals of the particular study, however, conversion to Cartesian distances may not be necessary, as the PC loads themselves are quite informative. This was not done in the current study because the PC loads served as satisfactory targets for neural network training, to which we now turn.
Part Two: Training a Neural Network to Estimate Articulation from Acoustics

METHODS

Network Architectures

A feedforward network (FFN; e.g., Rumelhart & McClelland, 1986) and a simple recurrent network (SRN; Elman, 1990) were trained to map acoustic input to articulatory output. Each network was comprised of 18 units in the input layer (one for each spectral bin), 8 units in the hidden layer, and 4 units in the output layer (one for each load on the first two lip PCs and the first two tongue PCs). In addition, the SRN had a layer of 8 context units. Adjacent layers were fully interconnected. See Figure 3.19.

![Neural network architectures: (left) feedforward network; (right) simple recurrent network. (Not all units are shown.)](image)

Activation Propagation

Input activations at each timestep were set to the values of the spectral bins for the

---

6In a pilot study SRNs and FFNs were tested with 8, 14, 20 and 30 hidden units. Generalization performance was best for both network types with 8 hidden units.
current audio frame. Standard activation propagation rules were used for each type of network, with logarithmic sigmoid units in the hidden and output layers. For the FFN, the activation, \( a_{in} \) of each hidden and output unit \( i \) was a function of its net input, \( net_i \), given by

\[
a_i = f(\text{net}_i) = f \left( \sum_j w_{ij} a_j + \text{bias}_i \right)
\]

where \( j \) ranges over the units which feed into unit \( i \), \( a_j \) is the activation of unit \( j \), \( w_{ij} \) is the weight of the connection to unit \( i \) from unit \( j \), \( \text{bias}_i \) is the bias term\(^7\) for unit \( i \), and \( f(\text{net}_i) \) is the logistic function

\[
f(\text{net}_i) = \frac{1}{1 + e^{-\text{net}_i}}.
\]

For the SRN, output unit activations were determined in the same manner as the feedforward network, but the activation of hidden unit \( i \) at time \( t \) was given by

\[
a_i = f(\text{net}_i) = f \left( \sum_j w_{ij} a_j(t) + \sum_k w_{ik} a_k(t) + \text{bias}_i \right)
\]

where, in addition to the previously defined symbols, \( k \) indexes the context units, \( w_{ik} \) is the weight on the connection from context unit \( k \) to hidden unit \( i \), and \( a_k(t) \) equals the activation of hidden unit \( k \) on the previous timestep, i.e., \( a_k(t) = a_k(t-1) \) for all \( i = k \).

Additionally, for the SRN, the hidden unit values were reset to zero at the start of each utterance.

**Training and Testing**

Three different data sets were created by using two of the three repetitions of the

---

\(^7\)The bias term may be thought of as an extra weighted connection to the unit from a unit that always has activation of 1.0.
corpus for training and reserving the other repetition for testing in a “jackknife” fashion. Each training set consisted of approximately 400 acoustic-articulatory pairs, or about 4.4 seconds of speech. Each test set consisted of about 200 pairs, or 2.2 seconds of speech. Three networks of each architecture type were trained on each dataset, for a total of 18 networks (3 datasets X 3 networks X 2 architectures).

At the start of training, connection weights were initialized with small random values. The network was then trained according to the backpropagation learning algorithm with momentum (Rumelhart et al., 1986). Specifically, weights were adjusted according to

\[ \Delta w_{ij}(t+1) = \eta \delta_i a_j + \alpha \Delta w_{ij}(t) \]

where the learning rate, \( \eta \), was set to 0.1, the momentum term, \( \alpha \), was set to 0.2, and error signal, \( \delta_i \), was calculated for output units according to

\[ \delta_i = (t_i - a_i) f'(net_i) \]

and for hidden units according to

\[ \delta_j = f'(net_j) \sum_i \delta_i w_{ij} \]

where \( f'(net) \) is the derivative of the activation function and \( t_i \) is the target activation for output unit \( i \).

Each network was trained for two million sweeps,\(^8\) where one sweep consisted of presentation of an acoustic input vector, propagation of activation to the output layer,

---

\(^8\)Pilot work determined that most networks approached an asymptote in the error curve (i.e., had learned most of what they would) by this point and risked overlearning if training was continued much longer. Overlearning is the phenomenon of learning the idiosyncratic details of the training set, causing performance on the test set to worsen. In fact, in some cases overlearning was seen before 2 million sweeps.
backward propagation of error, and weight update. This took about 30 min for each FFN and 90 min for each SRN using the tlearn simulator (Center for Research in Language, San Diego, CA) on an SGI Indigo 2 XL with a 100 MHz processor.

Each network was tested by sequentially presenting the acoustic input vectors of the test set (without error propagation or weight update) and recording output unit activations and errors. Output activations, which corresponded to the networks’ estimates of the principal component loads, were converted into images of the lip and tongue by mapping the range from 0.1-0.9 back to the original pre-normalized range and using the reconstruction procedure described in part 1. Movies were created from these images using the digital media conversion software dmconvert (Silicon Graphics, Inc., Mountain View, CA).

RESULTS AND DISCUSSION

Result #1: The SRN, but not the FFN, learned the most critical aspects of the acoustic-articulatory relationship.

Performance of the networks was judged both objectively, by looking at root mean square error (RMSE) and correlation (r) between actual and estimated PC load trajectories, and subjectively, by visually comparing the resulting video output. On both measures the SRNs consistently performed better than the FFNs. Summary results of all 18 trained networks are presented here first, followed by a detailed look at the best network of each architecture type.

In finding an objective measure of network performance on this task, it was important to consider more than just the standard measure of network error, RMSE. This is because a network may fail to learn the acoustic-articulatory relationship for an utterance but still reduce overall RMSE to a large extent by consistently predicting an average articulatory state — a response that clearly should not be counted as correct
(Papçun et al., 1992). A measure that captures whether the network is predicting the correct variations in the loads, and thus the correct articulatory movements, is the Pearson correlation \((r)\). Correlations are not a perfect measure on their own either, however, because in cases in which there is little articulatory motion, very small estimation errors may cause overall trajectories to have very low correlations. For this reason, both measures were analyzed, with less significance attributed to correlations when there were only small variations in the target loads. As discussed above, target variations were largest in the lip PCs during the /bəbəbə/ utterance and largest on tongue PC1 (which tracked back raising) and tongue PC2 (which tracked front raising) during the /ɡəɡəɡə/ and /dədədə/ utterances, respectively.

As shown in Table 3.1, the average correlations of the SRNs were much higher than that for the FFNs, especially for the PC loads with the greatest variation (bold). Table 3.2 shows that RMSE was also lower in most cases for the SRN, with greater average error for the SRN only on lip PC2 in the /dədədə/ and /ɡəɡəɡə/ utterances (although these differences did not reach the level of statistical significance as determined by t-tests). The high errors on these two PCs will be discussed in detail below.

In order to illustrate the types of errors common among the FFNs and the potential of the SRNs, the performance of the best network of each type is shown in Figure 3.20 and Figure 3.21, respectively. The FFN performed particularly poorly during the nearly silent periods associated with complete vocal tract closure, especially on lip PC1 and both tongue PCs. From the previous analysis of the various PCs, we know that lip PC1 tracks lip aperture, that tongue PC1 tracks back raising, and that tongue PC2 tracks front raising. Thus the FFN performed poorly on all three major articulatory
Table 3.1: Average correlation coefficients (with s.d.) between target and network output trajectories for the untrained test utterances of the nine networks of each type. Stars denote significant differences at the $p < 0.05$ level between that cell and the corresponding cell in the FFN. Scores for PCs with the highest target variations are shown in bold.

<table>
<thead>
<tr>
<th></th>
<th>FFN</th>
<th></th>
<th>SRN</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>/baba/</td>
<td>dada/</td>
<td>gaga/</td>
<td>/baba/</td>
<td>dada/</td>
</tr>
<tr>
<td>Lip PC1</td>
<td>0.20</td>
<td>0.58</td>
<td>0.19</td>
<td>0.90*</td>
<td>0.71*</td>
</tr>
<tr>
<td></td>
<td>(0.35)</td>
<td>(0.10)</td>
<td>(0.22)</td>
<td>(0.04)</td>
<td>(0.15)</td>
</tr>
<tr>
<td>Lip PC2</td>
<td>0.63</td>
<td>0.26</td>
<td>-0.17</td>
<td>0.86*</td>
<td>-0.02</td>
</tr>
<tr>
<td></td>
<td>(0.07)</td>
<td>(0.48)</td>
<td>(0.20)</td>
<td>(0.13)</td>
<td>(0.63)</td>
</tr>
<tr>
<td>Tng PC1</td>
<td>-0.29</td>
<td>0.25</td>
<td>-0.15</td>
<td>0.02</td>
<td>0.48*</td>
</tr>
<tr>
<td></td>
<td>(0.37)</td>
<td>(0.10)</td>
<td>(0.14)</td>
<td>(0.33)</td>
<td>(0.13)</td>
</tr>
<tr>
<td>Tng PC2</td>
<td>0.14</td>
<td>-0.05</td>
<td>0.14</td>
<td>0.19</td>
<td>0.55*</td>
</tr>
<tr>
<td></td>
<td>(0.35)</td>
<td>(0.23)</td>
<td>(0.36)</td>
<td>(0.39)</td>
<td>(0.24)</td>
</tr>
</tbody>
</table>

Table 3.2: Average root mean square error (with s.d.) on the untrained test utterances for the nine networks of each type. Stars denote significant differences at the $p < 0.05$ level between that cell and the corresponding cell in the FFN.

<table>
<thead>
<tr>
<th></th>
<th>FFN</th>
<th></th>
<th>SRN</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>/baba/</td>
<td>dada/</td>
<td>gaga/</td>
<td>/baba/</td>
<td>dada/</td>
</tr>
<tr>
<td>Lip PC1</td>
<td>0.21</td>
<td>0.17</td>
<td>0.16</td>
<td>0.14*</td>
<td>0.17</td>
</tr>
<tr>
<td></td>
<td>(0.03)</td>
<td>(0.09)</td>
<td>(0.06)</td>
<td>(0.05)</td>
<td>(0.09)</td>
</tr>
<tr>
<td>Lip PC2</td>
<td>0.23</td>
<td>0.26</td>
<td>0.18</td>
<td>0.15*</td>
<td>0.33</td>
</tr>
<tr>
<td></td>
<td>(0.02)</td>
<td>(0.13)</td>
<td>(0.14)</td>
<td>(0.05)</td>
<td>(0.15)</td>
</tr>
<tr>
<td>Tng PC1</td>
<td>0.11</td>
<td>0.11</td>
<td>0.16</td>
<td>0.06*</td>
<td>0.11</td>
</tr>
<tr>
<td></td>
<td>(0.04)</td>
<td>(0.03)</td>
<td>(0.03)</td>
<td>(0.02)</td>
<td>(0.04)</td>
</tr>
<tr>
<td>Tng PC2</td>
<td>0.13</td>
<td>0.25</td>
<td>0.15</td>
<td>0.10*</td>
<td>0.17</td>
</tr>
<tr>
<td></td>
<td>(0.02)</td>
<td>(0.07)</td>
<td>(0.03)</td>
<td>(0.03)</td>
<td>(0.14)</td>
</tr>
</tbody>
</table>
gestures of concern in this corpus. Video produced from reconstructions based on FFN output loads is consequently very poor in quality. Note, however, that the FFN did distinguish between the vowel portion of the three different utterances (at least on the two tongue PCs), indicating that the coarticulated schwas were acoustically distinct enough for the network to notice the difference.

Sample images of the lips from a /bɔbɔbɔ/ utterance are shown in Figure 3.22. Notice that the lip images reconstructed from the FFN output fail to produce the basic closed-open-closed pattern. In contrast, images produced from the SRN output loads
show a much better correspondence, although the lips do not open as wide as they should.

In movie format, tongue images reconstructed from the SRN output loads clearly showed the basic movement of each utterance, i.e., little movement for /bə/, front raising for /də/, and back raising for /gə/, even though the reconstructions of the tongue from only 2 PCs were too degraded to easily see the variation in a series of still images. (More PCs were used in subsequent studies.) An error in the timing of tongue tip motion for /dədədə/ is apparent when the SRN output movie is played in synchronization with the target movie, however. The PC load underpinnings of this timing error can be seen.
Figure 3.22: Sample articulatory estimates for /bab/ by the best FFN and best SRN. Images were reconstructed from the first two lip PC loads.

above in Figure 3.21. Tongue tip lowering was delayed by about 6 acoustic frames, or 60-70 ms.

The SRN is able to make better predictions than the FFN because of the SRN’s sensitivity to the dynamics of the utterances. Poor performance by the FFN is expected whenever there is an inconsistency in the mapping from acoustic input to articulatory output; that is, when many similar acoustic frames map to distinct articulatory images. In such cases, the FFN will tend to predict the average of the associated articulatory images. The SRN is able to disambiguate most of these inconsistent mappings because its estimates are based not only on the acoustic input, but also on the previous network state. For example, although the nearly silent periods in /bəbəbə/ may be acoustically similar to
the nearly silent periods in /gəɡəɡə/, the internal state of the network is different because of the recent acoustic differences in the vowels.

Result #2: Positions of non-critical articulators were not accurately estimated.

Non-critical articulators for a phoneme are those whose position causes little acoustic difference for that phoneme. As pointed out in part one of this chapter, the upper lip was raised in two of the three /dədədə/ and /ɡəɡəɡə/ utterances, causing a large change in lip PC2. An important question is whether this shift in the upper lip had acoustic consequences that would allow the estimation of upper lip height from acoustics. If not, then the upper lip may be a non-critical articulator for /də/ and /ɡə/.

As can be seen in Figure 3.20 and Figure 3.21 above, neither network architecture was able to produce good articulatory estimates of lip PC2 for these sounds. Analysis of the performance on the training samples showed that the basis for this error was indeed an inability to differentiate the acoustics of the articulatorily distinct /dədədə/ and /ɡəɡəɡə/ utterances. This inability is shown for the SRN on /dədədə/ in Figure 3.23. Notice that the SRN’s estimates of lip PC2 on the two training utterances are almost identical in spite of the considerable differences in the target values. The network’s response to the test utterance is also very similar. In informal listening tests, there were no auditorily perceptible differences in the three /dədədə/ utterances. These factors all indicate that the detected variation in upper lip position has little effect on the acoustics of /də/ and /ɡə/ and is therefore, within the detected range of variability, a non-critical articulator for these phonemes.

Fortunately, the very situation that creates this type of error also causes it to be of little importance in many inverse problem applications. The network fails here because the upper lip is not a critical articulator. But precisely because the upper lip is not a critical articulator, accurate estimation of its position is not critical for proper identification
Summary of Network Results

In summary, the static mapping from short acoustic frames to articulatory positions contained too many inconsistencies for the FFN to produce appropriate articulatory gestures. Performance was greatly enhanced by introducing recurrent connections that allowed the SRN to take into account dynamic aspects of the mapping. In particular, inconsistencies in the mapping for the nearly silent periods associated with consonantal closure did not prevent the SRN from estimating the proper articulatory gesture for each consonant. The SRN misestimated the timing of the tongue tip lowering gesture in /dₐdₐdₐ/ by about 60-70 ms, however. The other large error of the SRN, misestimation of lip PC2 during /dₐdₐdₐ/, appeared to be due to a variation in articulation.
that had little acoustic consequence, a slight raising of the upper lip. Non-critical articulators such as this may defy estimation from acoustics, but such errors may also be of little importance in some applications because of their non-critical status.

This study was of course limited in some ways. With only three repetitions of each utterance, it is hard to know how robust these findings are. Also, as discussed above, the appearance of the white teeth in the otherwise dark oral cavity may have exaggerated the variability in lip aperture. The appearance of the teeth also sent a mixed signal — with no lip opening the pixels at the center of the mouth are fairly bright (representing the lips), and with moderate opening the pixels are dark (the oral cavity), and then with even more opening the pixels are light again (the teeth) — thus making the relationship between extracted PCs and lip aperture more complex than it need be. The next chapter describes changes to the method that were implemented in order to address these limitations and otherwise improve performance.
CHAPTER 4: IMPROVING THE METHOD — LONG WINDOWS AND PAINTED FACES

This chapter is principally about fine-tuning the acoustic-articulatory mapping method described in Chapter 3. Changes were made in speech materials, recording methods, data processing, network training, and image reconstruction techniques, and served to confirm and clarify the results presented in Chapter 3. Because many of the changes in recording and data processing methods were motivated by concerns about network training, data processing and neural network methods are presented this time in a single methods section. Results are presented in a format that runs parallel to Chapter 3, however, with PCA results presented first, followed by the results of network training.

METHODS

The methods used were the same as those presented in Chapter 3 with the following changes.

Data Collection: More repetitions and cleaner images

The speech corpus consisted of the same voiced stops but this time embedded in a schwa on either side, i.e., /zβaʊ əʊ əʊ/, /dθəʊ əʊ əʊ/, /θəʊ əʊ əʊ/. This had the effect of slightly reducing coarticulation and increasing movement. Each utterance was spoken without internal pause and with slight emphasis on the second schwa in each triple. A natural speaking rate was used with each utterance about 1.8 sec in duration. Importantly, the set of three utterances was spoken ten times rather than three in order to increase the number of samples for network training.

Ultrasound recordings were as before except that the focal depth and gain of the transducer were carefully set so as to remove as much of the reflectance as possible from
the internal structures of the tongue. The surface contours were thus captured more clearly.

For the video recording, in order to remove the complexity added by the appearance of the white teeth in a dark oral cavity, the teeth were blackened with a dyed wax. The lips were also painted white and the surrounding area of the face black with theatrical make-up. This served to clarify the outline of the lips to make precise lip movements more detectable.

**Video Processing: Blackening of uninformative regions**

The video and ultrasound signals were processed just as before except that certain corner portions of the rectangles containing the lip and tongue images were blackened by setting the pixel intensities to zero. This was necessary for the tongue images because the video image of the lips was inadvertently inserted with the video mixer too close to the tongue image to allow cropping of the entire tongue image without also including some of the video camera image. It was decided also to blacken corners in the lip images simply to reduce pixel intensity variations in these non-lip-containing areas of the images.

**Acoustic Processing: Three acoustic conditions**

For acoustics, in addition to using a 22.2-ms window and frame rate of 11.1 ms, an additional set of acoustic data was acquired in which a 44.4-ms window was used with a 33.3-ms frame rate (thus also overlapping by 11.1 ms). In network training these two types of audio processing were referred to as the “short window” and “long window” conditions. Note that the long window condition did not require the interpolation of loads between video frames captured at the standard video frame rate. Lengthening of the acoustic analysis window was thought to offer the potential advantage of giving the network a broader view of the signal on which to base an articulatory estimate, albeit at
the risk of blurring acoustic changes within that window. In order to further broaden the “acoustic view” without excessively blurring changes over time, a third condition was also used in which 3 long (44.4-ms) windows were presented to the network simultaneously, with the middle acoustic window aligned with the current (target) video frame. The total width of the acoustic window presented to the networks in this “3 long windows” condition was 111 ms, with a new set of three windows taken every 33.3 ms.

**Network Architectures: Larger networks to estimate more PCs**

As in the previous study, the networks had 18 input and 8 hidden units, except for the networks trained in the 3 long windows condition in which 54 input units (3 windows × 18 bins) were required and 12 hidden units were found to produce good generalization.¹ The number of output units was determined by analysis of the PC loads and images reconstructed therefrom. Because the sixth tongue PC seemed to be importantly involved in tracking tongue raising, and because reasonably good quality images of both the lips and the tongue could be reconstructed from six PCs, six output units were used, one for each PC.

**Network Training: Separation of the lip and tongue tasks**

Unlike the previous study, separate networks were trained for lip and tongue images. Because there was little correlation between lip and tongue movement in this speech corpus, separation of the tasks was thought to make each task easier. This also kept the number of output units small (6 instead of 12), thus reducing network size and

---

¹Trials to determine an appropriate number of hidden units were conducted with 4, 8, 15 and 30 hidden units for the networks with 18 input units, and with 6, 12, and 24 hidden units for the networks with 54 input units.
training times. Because the networks received identical input, separately trained lip and
tongue networks could be combined to produce a composite estimate of overall vocal
tract shape if that were deemed important.

Networks were trained on nine of the corpus repetitions and tested on the tenth.
Five networks of each architectural type (SRN and FFN) and each articulatory target (lips
and tongue), were trained under each acoustic condition (short window, long window,
and 3 long windows), each time reserving a different utterance for testing. This resulted
in a total of 60 trained networks. With more and longer utterances, the size of the training
sets were much larger than in the previous study. In the short windows condition there
were about 4500 acoustic-articulatory pairs in each training set and about 170 pairs in
each test set, with one-third this number in each of the other two acoustic conditions.
Networks were trained for 2,500,000 sweeps. Training times varied from about half and
hour to about three hours per network depending on network size.

Network Testing: Using output loads as an index to images in the training corpus

In addition to using the network output PC loads to reconstruct lip and tongue
images directly as before, network output movies were also created by selecting, for each
set of network outputs, the image from (a subset of) the training images with the most
similar PC loads, as defined by Euclidean distance. Each frame of a network output
movie was thus an original, sharp image without the blurred edges sometimes found in
reconstructions. Such movies were called “nearest neighbor movies” to distinguish them
from the “reconstructed movies.” Although each frame of a nearest neighbor movie was
crisp, frame to frame differences were generally larger than that for reconstructed movies,
sometimes causing the movies to seem less smooth, especially if network errors were
large.
In order to keep the nearest neighbor search in six-dimensional PC load space computationally tractable, only a subset of the approximately 1500 training images were used as candidate nearest neighbors. This subset of images, called the “spanning set,” was selected by a procedure designed to span the full range of image types with a minimal number of images (in essence, to throw out duplicates). The procedure went as follows: the first image was added to the spanning set; each consecutive image in the corpus was then added to the spanning set only if it was at least Euclidean distance, $d$, in PC load space from any other image already in the spanning set. A distance of $d = 0.1$ was found empirically to provide a relatively small number of images (213 for the lips, 452 for the tongue) from which smooth nearest neighbor movies could be made.

**PCA Results and Discussion**

PCA again produced representations of the lips and tongue useful not only for network training but also for speech production analysis. As in the previous study, analysis of the PC loads provided clear indication of coarticulation, temporal asymmetries, and production variability.

Sample images indicating the additional clarity of the new recording technique are show in Figure 4.1, the average lip and tongue images in Figure 4.2, and the first several PCs in Figure 4.3. Although the first few lip PCs explained more of the variance in the dataset than those for the unpainted face in the previous study, less variance was explained than previously by the first several tongue PCs (Figure 4.4). It is not clear, however, whether there is actually less systematicity in this set of tongue images or whether the decrease in variance explained by the tongue PCs is simply due to the “snow” in the now eight times larger number of ultrasound images.
Figure 4.1: Sample images from the three utterance types. Consecutive frames are shown from complete consonantal closure to the point of maximal opening.

Figure 4.2: The average lip and average tongue images.

Figure 4.3: The first seven principal components of the lips and tongue scaled in intensity to indicate relative importance.
Figure 4.4: The percent variance explained by the first 20 PCs of the lip and tongue images. The solid line indicates the cumulative percent variance explained.

A good indicator of the role of each of the PCs in this dataset is a plot of the value of the loads on the PCs over the course of the utterances. Such a plot is shown for the lip PCs in Figure 4.5. Lip PC1 follows the basic open-closed-open pattern of the lips in the /əbə əbə əbə/ utterances. Lip PC2 reaches a maximum, and lip PC4 a minimum, midway through closure and opening, thus indicating a partially open mouth. Lip PC5 and lip PC6 show a marked asymmetry over the course of an /əbə/ gesture, at one extreme as the lips move toward closure and at the other as the lips move towards complete opening. This is the same temporal asymmetry discovered previously, but it manifests itself much more clearly in these images. As Figure 4.6 shows, the lips open first at the midline but close more evenly.
Another aspect of speech production made clear in Figure 4.5 is the variability in lip aperture across the /ædədədə/ utterance. Notice in particular the variation in lip PC2. The physical manifestation of this variability is shown in Figure 4.7. The lips were open wider in the first than in the third /ædə/ by about 4 mm. Comparisons across the ten utterances of the same type revealed a similar variability of about 4 mm in lip aperture for /ægəægəægə/. 

Figure 4.5: Variation in lip PC loads over time for three sample utterances.
Figure 4.6: Differences in lip motion during the opening and closing of /aβa/. The closure sequence is reversed to facilitate comparison. The lips open first at the midline, but close more evenly.

Figure 4.7: Variability in the production of /aβa/. Frames are taken from the first and third /aβa/ repetitions in the /aβa aβa aβa/ utterance of Figure 4.5.
A parallel analysis of the tongue PC loads over time revealed that the tongue front raising gesture of /aða/ was mostly strongly tracked by tongue PCs 3 and 6, and that the back raising gesture of /aɡa/ was most strongly tracked by tongue PCs 1, 2, and 4 (Figure 4.8). When one front raising component is plotted versus one back raising component, the consonantal coloring of the schwa in /aða aða aða/ and /aɡa aɡa aɡa/ is indicated by the lack of overlap (Figure 4.9).

![Variation in tongue PC loads over time for three sample utterances.](image)

**Summary of PCA Results**

PCA of raw lip and ultrasound images proved once again revealing of subtle aspects of speech production. PC load analysis showed clear temporal asymmetries in lip
movement and coarticulatory influences on the tongue. Variability in lip aperture for non-critical articulators was also found, providing another opportunity to test the ability of the networks to estimate the position of non-critical articulators from acoustics.

**NEURAL NETWORK RESULTS AND DISCUSSION**

**Result #1: Longer acoustic analysis windows lead to better performance.**

The long windows condition (44.4-ms acoustic windows with a 33.3-ms frame rate) improved performance over the short windows (22.2-ms acoustic windows with an 11.1-ms frame rate) condition for both the SRN and the FFN. Performance was again judged by two criteria: root mean square error and correlation between output and target trajectories. Because correlations are less meaningful when there is little variation in the target load (as discussed in Chapter 3), average correlation coefficients were calculated.
over only those PCs whose loads showed significant variation for each utterance type, i.e., lip PCs 1 and 2 for /ɒ ʌ ə/ , tongue PCs 3 and 6 for /æ ə æ ə/ , and tongue PCs 1 and 2 for /æɡ əɡ əɡ/ . Table 4.1 shows the average correlations over these PCs for each network type and each acoustic condition. There are five networks trained on the lips and five trained on the tongue for a total of ten networks in each cell. Stars denote a significant difference at the $p < 0.05$ level between that cell and the cell immediately above as determined by t-tests. Table 4.2 shows the root mean square error (over all PCs for all utterance types) in the same format. Although both network types showed an average reduction in error with the change to longer acoustic windows, the differences just failed to meet the $p < 0.05$ level of statistical significance. With the corresponding significant improvement in correlations, however, we can confidently conclude that the longer windows improved overall performance.

The change from the long window condition to the 3 long windows condition (in which three acoustic windows were presented simultaneously at each timestep) improved the performance of the FFN, but had no significant effect on the SRN. The single best network of each architecture type for overall performance was found in the 3 long windows condition, however.

Table 4.1: Average correlation coefficients (with s.d.) between target and network output trajectories for the most critical PCs of each test utterance (see text). Stars denote a significant difference at the $p = .05$ level between that cell and the cell immediately above.

<table>
<thead>
<tr>
<th>Acoustic Condition</th>
<th>FFN</th>
<th>SRN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Short Window</td>
<td>0.71</td>
<td>0.75</td>
</tr>
<tr>
<td></td>
<td>(0.11)</td>
<td>(0.14)</td>
</tr>
<tr>
<td>Long Window</td>
<td>0.82*</td>
<td>0.86*</td>
</tr>
<tr>
<td></td>
<td>(0.08)</td>
<td>(0.05)</td>
</tr>
<tr>
<td>3 Long Windows</td>
<td>0.91*</td>
<td>0.88</td>
</tr>
<tr>
<td></td>
<td>(0.04)</td>
<td>(0.10)</td>
</tr>
</tbody>
</table>
Table 4.2: Average root mean square error (with s.d.) on the test utterances.

<table>
<thead>
<tr>
<th>Acoustic Condition</th>
<th>FFN</th>
<th>SRN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Short Window</td>
<td>0.14</td>
<td>0.13</td>
</tr>
<tr>
<td></td>
<td>(0.03)</td>
<td>(0.03)</td>
</tr>
<tr>
<td>Long Window</td>
<td>0.12</td>
<td>0.10</td>
</tr>
<tr>
<td></td>
<td>(0.03)</td>
<td>(0.03)</td>
</tr>
<tr>
<td>3 Long Windows</td>
<td>0.11</td>
<td>0.11</td>
</tr>
<tr>
<td></td>
<td>(0.02)</td>
<td>(0.02)</td>
</tr>
</tbody>
</table>

Result #2: Only the SRN learned the critical aspects of the acoustic-articulatory relationship.

In overall comparison of the FFNs and SRNs as presented in Tables 4.1 and 4.2, there were no statistically significant differences between the two architectural types. Indeed, looking at just these overall comparisons it appears that the FFNs might have even performed better than the SRNs in the 3 long windows condition. A closer look at the individual performance of each network revealed, however, that the FFNs reduced overall error without actually learning to distinguish between the three utterance types. This was possible because such a large percentage of the training pairs in this corpus represented a vowel produced with a relatively open vocal tract. For example, the only short window FFN to produce lip closure for /əbə əbə əbə/ also produced lip closure during the stops of the non-labial utterances. Conversely, the short window FFNs which did not close the lips at /d/ and /g/ failed to close the lips during /b/. In contrast, three of the five short window SRNs produced appropriate lip closure for all three test utterances.

The situation was much the same for the tongue. For the most part, networks that produced a smooth and convincing back raising for /g/ failed to produce the appropriate front raising for /d/ and vice versa. The only networks that were able to produce the appropriate gestures for both /d/ and /g/ were all SRNs. Figure 4.10 illustrates this point.
for all 30 networks trained on the tongue images. This figure shows that for the primary component of back raising, PC1, and for the primary component of front raising, PC3, the networks tended to make small errors on /adə adə ada/ or /əgə aəgə aəgə/ but not both. SRNs are represented by Xs and FFNs by Os. The few networks to do well on both utterances were all SRNs. The four SRNs that performed well on both PCs in both utterances are shown in bold. Of these, two were trained in the long window condition and two in the 3 long windows condition.

Although the output is best appreciated in movie format, sample images from the best FFNs and best SRNs are shown for the lips for /aba/ and for the tongue for /adə/ and /əgə/ in Figures 4.11 – 4.13. In each case, reconstruction images (images made by recombining the PCs according the values of the network output loads) are shown first, followed by nearest neighbor images (images with most similar loads from the spanning set of the training corpus). The nearest neighbor images are clearer in still format but sometimes choppy in movie format. Notice that both networks trained on the lips produced appropriate labial gestures for all three test utterances. The SRN trained on the
Figure 4.11: Lip images from /aba/ for the best FFN and the best SRN. Images in the top block were reconstructed from the output loads. Images in the bottom block were found by the nearest neighbor method.
Figure 4.12: Tongue images from /ææ/ for the best FFN and the best SRN. Images in the top block were reconstructed from the output loads. Images in the bottom block were found by the nearest neighbor method.
Figure 4.13: Tongue images from /ægæ/ for the best FFN and the best SRN. Images in the top block were reconstructed from the output loads. Images in the bottom block were found by the nearest neighbor method.
tongue produced an appropriate and smooth gesture for all three test utterances, but the FFN trained on the tongue performed better on /d/ than on /g/ (some other FFNs showed the opposite pattern).

**Result #3:** Estimates of a non-critical articulator tended toward the average for that articulatory gesture, in a sense “correcting” unusual articulation.

As mentioned above, the amount of lip aperture varied in the production of the non-labial sounds. For example, lip aperture in the sixth repetition of /əɡə əɡə əɡə/ was unusually small (Figure 4.14). None of the networks were able to find a difference in acoustics that allowed them to produce an appropriately adjusted estimate of lip aperture for this utterance. The variation in lip aperture was most strongly registered in lip PCs 1 and 2. For the sixth repetition of /əɡə əɡə əɡə/, lip PC1 was unusually low and lip PC2 was unusually high. Figure 4.15 shows the range of estimates of these PCs over the sixth /əɡə əɡə əɡə/ repetition for the 24 networks in which this repetition was a part of the training set. (The other six networks had no chance to learn the variation as they were not exposed to it in training.) In every case but one, the estimates consistently fell short of the target range in the expected direction (too high for PC1 and too low for PC2). The one exception was an FFN which predicted complete lip closure for all stops (i.e., it failed even worse than the other 23 networks).

Notice that the failure to accurately estimate lip position in this case does not produce a lip image that is inappropriate for an /əɡə/ utterance, however. On the contrary, the error results in a lip image that looks more like a typical /əɡə/ lip gesture than the one actually used. Thus, what is an error from the point of view of strictly accurate articulatory estimation from speech sounds can be seen as a “correction” of unusual articulation if one's goal is to produce a canonical articulatory representation. A canonical representation may be as good as (or possibly even superior to) an authentic
Figure 4.14: Lip images from the sixth and tenth repetitions of the /ægæægæ/ utterance. Lip aperture in the sixth repetition was unusually small.

Figure 4.15: Failure to estimate exceptionally small lip aperture in the sixth repetition of /ægæægæ/. The dotted line shows the average estimate of the indicated PC for each of the 24 networks trained on the sixth repetition. The solid line indicates the average of the actual value for the PC. Error bars represent one standard deviation.
representation for some applications. This is particularly true for the production of speechreadable articulatory gestures because it tends to associate meaningful acoustic units with standard, and thus more easily learnable, articulatory shapes. In effect, it may hide confusing idiosyncrasies.

**Summary of Network Results**

Network performance under these revised conditions once again supported the notion that the most critical aspects of articulation are estimable from speech acoustics, at least for a small corpus of voiced stops in a schwa context. The sensitivity to temporal change afforded by the recurrent connections of the SRN proved important in learning to distinguish the three gesture types. Longer (44.4-ms) windows and the use of three simultaneously presented windows also improved performance over the acoustic analysis method used in Chapter 3. Variability in non-critical articulators was not learned by the networks, but this failure did not prevent the networks from producing appropriate canonical vocal tract shapes.
CHAPTER 5: GENERATING SPEECHREADABLE VOCAL TRACT IMAGES FROM ACOUSTIC INPUT

The previous chapters have established that articulation can be estimated from acoustics for a simple corpus composed of four phonemes. This chapter presents the study of a more complex corpus composed of seventeen different phonemes, including the word “say” and the spoken digits 0-7. In addition to the testing methods used before, network performance on this larger corpus was evaluated by presenting the movies derived from network output to human subjects for speechreading.

It was hypothesized that the vocal tract images produced by the network could be used to improve the speech comprehension of persons with hearing loss in situations in which the speech signal could reasonably be sent to a computer with a video display, such as on the telephone or in a classroom or business meeting. In order to provide a first test of this hypothesis, normal hearing volunteers were provided with the network output movies together with a degraded audio signal designed to simulate hearing impairment. Simulation of hearing loss has been successfully achieved both by frequency-specific attenuation (filtering) and by masking with a shaped masking noise (Fabry & Van Tasell, 1986). This study used extreme low-pass filtering to simulate the experience of the severely, but not totally, hearing impaired, the population for which the device was thought likely to be of most use. The results show that this method holds much promise as a special situations assistive listening device.

METHODS

Speech Materials

The speech corpus consisted of 24 different utterances spoken by the author six times each and recorded in same manner as described in Chapter 4. Each utterance was
composed of the word “say” followed by a series of three digits from the set 0-7, with 0 pronounced “oh”; the corpus was named the “numbers” speech corpus. Digit sequences were created by concatenating three randomly ordered lists composed of three repetitions of each digit, so that each digit appeared three times in each position. The set of digit sequences used is listed in Table 5.1. Utterances were spoken at a comfortable rate, with the average duration of an utterance about 2.5 sec.

Table 5.1: Digit sequences used in the speech corpus. Each sequence was preceded by the word “say.” Stars indicate sequences reserved from the network training set.

<table>
<thead>
<tr>
<th>011</th>
<th>137*</th>
<th>215</th>
<th>322</th>
<th>431*</th>
<th>501*</th>
<th>600</th>
<th>703</th>
</tr>
</thead>
<tbody>
<tr>
<td>024</td>
<td>162</td>
<td>254*</td>
<td>324*</td>
<td>437</td>
<td>510</td>
<td>640*</td>
<td>752*</td>
</tr>
<tr>
<td>045*</td>
<td>175</td>
<td>263</td>
<td>376</td>
<td>456</td>
<td>547</td>
<td>673</td>
<td>766</td>
</tr>
</tbody>
</table>

Data Processing and Network Training

Data was processed and networks were trained according to the procedures described in Chapter 4 with the following exceptions. Only the “3 long windows” acoustic processing condition was used. Also, more principal components (eight instead of six) were used for each articulator because the speech corpus included a wider variety of lip and tongue shapes than before so that each of the PCs captured a smaller percentage of the variance. Separate networks with 30 hidden units each were trained to estimate each individual PC, and then combined into a large composite network for testing. One composite SRN and one composite FFN were created for each articulator, for a total of 4 composite networks. This “divide and conquer” technique was thought to make the task of each network easier, thereby improving performance. Pilot study comparisons of single-output networks to networks trained on all eight PCs simultaneously indicated that this was indeed the case.
In order to further test the generalization ability of the networks, eight of the recorded utterances (starred in Table 5.1) were reserved from the training set such that half of the digit pairs in the reserved set did not appear in the same order in the training set. These reserved utterances were called “novel” utterances (though all utterances on which the network was tested were novel in the sense that the network had not been trained on that particular token). The networks were trained on four repetitions of the remaining 16 (“familiar”) utterance types for a total of 64 utterance tokens, or 4758 acoustic-articulatory pairs. This represented about 2.6 minutes of speech. The networks were tested on a novel (untrained) token of all 24 utterance types (1723 pairs, or about 1 minute of speech). Networks were trained up to 5 million sweeps, with weights saved every 1 million sweeps so that weight values could be discarded if overlearning was encountered. Each network exhibited its best performance from 3 million to 5 million sweeps. Training times ranged from 5 to 18 hours for each composite network on a 450 MHz processor.

“Speechreadability” Testing

The “speechreadability” of the SRN output movies of the lips was tested by presenting them for recognition to ten normal hearing volunteers. The movies were scaled to approximate the size of normal adult lips and displayed on the computer monitor. Subjects were asked to type the three-digit sequences contained in utterances presented in three different conditions. In the video (V) condition subjects saw the movie of the lips only. In the audio (A) condition, subjects saw a blank white screen and heard through earphones an audio signal which had been low-pass filtered with a seventh order Butterworth filter with a cutoff frequency of 100 Hz (Figure 5.1). In the audio-visual (AV) condition, subjects saw the moving lips presented synchronously with the degraded audio signal. In each condition containing a video signal, movies generated from the network output were interspersed with originally recorded movies. All network output
movies were created using the nearest neighbor method described in Chapter 4. (For this corpus, a distance of \( d = 0.1 \) generated a spanning set of 857 images from which images could be drawn.)

![Fig. 5.1: The effect of low-pass filtering at 100 Hz as shown by the spectrogram of the utterance “say five four seven.”](image)

Subjects were informed that all stimuli would contain the word “say” followed by a sequence of three digits from 0-7, with 0 pronounced “oh,” and repetition of digits possible. Subjects were informed about the degraded nature of the audio signal, and about the possible difficulty of lipreading, and asked to guess when uncertain. Subjects were tested individually and controlled the pace of the experiment.

At the beginning of the experiment subjects were given three practice blocks of four trials each (one block in each condition) with feedback. These practice trials were recorded but not scored. At the end of practice, subjects were warned that they would no longer receive feedback. Subjects were alerted at the beginning of each block whether they would be presented with A, V, or AV. In the test phase, five blocks of 12 stimuli each were presented in the order AV, V, A, V, AV. This order was used to minimize possible practice or fatigue effects on any one condition. Post hoc analysis of performance on the first 30 vs. the last 30 non-practice trials showed no change in performance, however. In each block with a video signal, half of the items were original
video and half were created from network output (unbeknownst to the subject). The set of 24 test utterances was cycled through in a random order three times during the experiment, subject to the following constraints: immediate repetition of an item was not allowed; exactly four of the novel utterances were included in each sequence of 12 utterances so that the novel and familiar utterances would be equally spread through the experimental conditions. Under these constraints it was not possible to prevent the presentation of some utterances more than once in the same format, but post hoc analysis showed that performance was unchanged by repeated presentation.

RESULTS AND DISCUSSION

The Images and Their Principal Components

Figure 5.2 shows the greater diversity of lip and tongue images obtained in this corpus as compared to the previous corpora. The average images and eigenimages are consequently more smoothed in appearance (Figures 5.3 and 5.4), with each of the first several PCs capturing less variation than before (Figure 5.5). The first eight PCs of the lips captured 90.3% of the variance and the first eight tongue PCs, 45.1%.

Figure 5.2: The 24 most distinct lip and tongue images from the “numbers” speech corpus.
Figure 5.3: The average lip and average tongue images.

Figure 5.4: The first eight principal components of the lips and tongue scaled in intensity to indicate relative importance.

Figure 5.5: The percent variance explained by the first 20 PCs of the lip and tongue images. The solid line indicates the cumulative percent variance explained.

**Network Performance**

The values of these PCs were successfully associated with variations in the acoustic signal by both neural network architectures. Average measures of root mean square error (RMSE) and correlation (CORR) between output and target trajectories are
shown in Table 5.2. Overall RMSE was actually somewhat lower for this phonemically more complex corpus than for the corpora in the previous studies. Average correlations for the individual PCs over all the test utterances were also high, ranging from 0.58 to 0.94. Correlational averages over all PCs for all test utterances are reported in Table 5.2. (Although the grand average correlations for this corpus were much higher than the grand average correlations for the previously studied corpora, direct comparison of these correlations is inappropriate due to the un informativeness of correlational values for the PCs with a small range of variation in the previous studies. In the phonemically diverse utterances of the “numbers” corpus, most PCs tended to show a large range of variation over most utterances.)

Table 5.2: Average root mean square error (with s.d.) and correlation coefficients (with s.d.) between target and network output trajectories for all eight PCs of each composite network. Scores are for the (untrained) test utterances only.

<table>
<thead>
<tr>
<th></th>
<th>RMSE</th>
<th>CORR</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FFN</td>
<td>SRN</td>
</tr>
<tr>
<td>Lips</td>
<td>0.10</td>
<td>0.10</td>
</tr>
<tr>
<td></td>
<td>(0.03)</td>
<td>(0.03)</td>
</tr>
<tr>
<td>Tongue</td>
<td>0.08</td>
<td>0.08</td>
</tr>
<tr>
<td></td>
<td>(0.01)</td>
<td>(0.02)</td>
</tr>
</tbody>
</table>

Comparisons of the performance of the two architectural types showed no significant differences. In fact, the SRN and FFN tended to make very much the same estimation errors — the output of the networks was even more similar to each other than it was to the target as judged by root mean square differences and trajectory correlations. One possible reason for this unexpected finding is that the task required less attention to temporal information than in the previous studies. In those studies, the principal distinguishing feature between the different utterances was a signal at the phoneme transitions, especially a rise vs. a fall in the second formant, which the SRN could benefit
from “remembering” in the activation values of its context units. The acoustic-articulatory relationships in the “numbers” corpus seemed to be less dependent on such temporal information, and may have therefore put less pressure for learning on the weights from the context units. Although the values of those weights were not extraordinarily small (they were in the same range as weights in other parts of the network), scrambling the presentation order of the acoustic input (and then “unscrambling” the output order) did not affect performance as severely as one would expect if the recurrent connections were playing an important role. On lip PC1, for example, RMSE rose from 0.13 to 0.15 and the average correlation score dropped from 0.88 to 0.42, indicating that performance was worse, but not catastrophically so. This indicates that the two network types may solve a problem in very much the same way if there is little to be gained by paying attention to dynamics, or, possibly, if the useful dynamical information is hard to extract. Because the network outputs were so similar, the output from only one of the networks (the SRN) was used in the speechreading test.

**Speechreadability of the Network Output Movies**

When no audio signal was available, correct spoken digit recognition was 59.4% with network video and 84.2% with the original video. In other words, the network video produced a level of intelligibility 70% as great as that produced by original lip images and almost five times greater than chance (12.5%). See Figure 5.6.

Under conditions of simulated hearing impairment, spoken digit recognition was significantly improved from 46.7% in the degraded audio condition to 80.3% when the audio signal was supplemented with the video produced by the neural network ($F = 8, p = 6.5 \times 10^{-9}$). Taking the scores produced in the audio plus original video condition
(93.6% correct) as the theoretical limit on intelligibility by these means, the audio plus network video condition realized 86% of its potential intelligibility score. See Figure 5.7.

The network’s ability to generalize to novel utterances was tested by comparing digit recognition performance on familiar and novel stimuli. (The term “novel” is used here to indicate digit sequences that the network was not exposed to during training; even the “familiar” stimuli were novel in the sense that the network had never seen the particular token on which it was tested.) Performance on familiar vs. novel stimuli with and without the degraded audio signal is presented in Table 5.3. As expected, performance was somewhat better on the familiar utterances, though not significantly so in the video only condition. Importantly, the network output even for novel utterances improved recognition from 46.7% (degraded audio alone) to 69.1% (with novel video), and this improvement was statistically significant \((F = 8, p = .002)\).
Figure 5.7: Percent correct recognition of spoken digits under conditions of simulated hearing impairment. Error bars represent one standard deviation.

Of course this test is preliminary in many respects — it is speaker dependent, limited vocabulary, and tested under forced-choice conditions (but without the aid of a meaningful context) — and claims only to suggest rather than prove the potential utility of such a system.

**Analysis of Network Errors and Suggestions for Improvement**

By analyzing confusion matrices we can determine which digits were most affected by which experimental conditions. Of particular interest is where the network output video failed, that is, where the network video induced errors that were not made when viewing the original video. Confusion matrices for the conditions that did not involve the neural network are presented in Table 5.4 and those for the neural network video are presented in Table 5.5.
Table 5.3: Percent correct spoken digit recognition (with s.d.) for familiar vs. novel utterances. Stars denote a significant difference at the $p < .05$ level from the cell immediately above.

<table>
<thead>
<tr>
<th></th>
<th>Video Only</th>
<th>Video with Degraded Audio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Novel</td>
<td>58.8 (19.3)</td>
<td>69.1 (18.7)</td>
</tr>
<tr>
<td>Familiar</td>
<td>60.6 (11.4)</td>
<td>85.5* (8.7)</td>
</tr>
</tbody>
</table>

Analysis of percent correct recognition for each digit (bold) indicates that performance gains from the network images followed closely the gains from the original video. Notice first that performance on “six” was particularly poor for the network video but also relatively poor for the original video. (When the audio signal was included, “six” was easily recognized.) Second, in the network video “two” and “oh” were highly confused, but the same is true of the original video. Third, performance was substantially improved over audio alone by both network and original video for every stimulus type not already at ceiling. These three results taken together indicate that the network video improved comprehension across the board and in much the same way as did the original video, only to a lesser degree.

There were, however, two patterns in the network video responses that were notably different from patterns of responses to the original video. First, the network video, with or without audio, elicited a large number of false “three” responses (especially for the stimuli “one,” “four,” and “five”). Although there were also a large number of false “three” responses in the audio only condition, which may indicate a simple bias for that response in the face of uncertainty, analysis of the network video output found a number of erroneous frames in which the tongue faintly protruded between the lips. As this is a strong visual indication of the phoneme /θ/, which was only present in “three,” false positives for this digit would be expected. Many of the
Table 5.4: Confusion matrices for the audio, visual, and audio-visual conditions in which only original video was presented. Values represent the percent of responses for each stimulus. Correct responses are in bold.

### Degraded Audio

<table>
<thead>
<tr>
<th>Stimulus</th>
<th>oh</th>
<th>one</th>
<th>two</th>
<th>three</th>
<th>four</th>
<th>five</th>
<th>six</th>
<th>seven</th>
</tr>
</thead>
<tbody>
<tr>
<td>oh</td>
<td>9</td>
<td>9</td>
<td>11</td>
<td>37</td>
<td>7</td>
<td>15</td>
<td>0</td>
<td>13</td>
</tr>
<tr>
<td>one</td>
<td>21</td>
<td>50</td>
<td>2</td>
<td>12</td>
<td>0</td>
<td>6</td>
<td>0</td>
<td>8</td>
</tr>
<tr>
<td>two</td>
<td>0</td>
<td>0</td>
<td>62</td>
<td>26</td>
<td>5</td>
<td>5</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>three</td>
<td>9</td>
<td>3</td>
<td>0</td>
<td>67</td>
<td>9</td>
<td>9</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>four</td>
<td>2</td>
<td>0</td>
<td>22</td>
<td>30</td>
<td>8</td>
<td>4</td>
<td>0</td>
<td>34</td>
</tr>
<tr>
<td>five</td>
<td>4</td>
<td>4</td>
<td>11</td>
<td>37</td>
<td>4</td>
<td>15</td>
<td>0</td>
<td>26</td>
</tr>
<tr>
<td>six</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>98</td>
<td>0</td>
</tr>
<tr>
<td>seven</td>
<td>5</td>
<td>0</td>
<td>2</td>
<td>7</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>83</td>
</tr>
</tbody>
</table>

### Original Video

<table>
<thead>
<tr>
<th>Stimulus</th>
<th>oh</th>
<th>one</th>
<th>two</th>
<th>three</th>
<th>four</th>
<th>five</th>
<th>six</th>
<th>seven</th>
</tr>
</thead>
<tbody>
<tr>
<td>oh</td>
<td>96</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>one</td>
<td>4</td>
<td>84</td>
<td>2</td>
<td>4</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>two</td>
<td>36</td>
<td>2</td>
<td>55</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>three</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>96</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>four</td>
<td>3</td>
<td>3</td>
<td>0</td>
<td>95</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>five</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>100</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>six</td>
<td>12</td>
<td>20</td>
<td>4</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>57</td>
<td>4</td>
</tr>
<tr>
<td>seven</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>96</td>
</tr>
</tbody>
</table>

### Degraded Audio with Original Video

<table>
<thead>
<tr>
<th>Stimulus</th>
<th>oh</th>
<th>one</th>
<th>two</th>
<th>three</th>
<th>four</th>
<th>five</th>
<th>six</th>
<th>seven</th>
</tr>
</thead>
<tbody>
<tr>
<td>oh</td>
<td>73</td>
<td>0</td>
<td>24</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>one</td>
<td>0</td>
<td>86</td>
<td>0</td>
<td>14</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>two</td>
<td>0</td>
<td>0</td>
<td>100</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>three</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>100</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>four</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>95</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>five</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>98</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>six</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>100</td>
<td>0</td>
</tr>
<tr>
<td>seven</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>100</td>
</tr>
</tbody>
</table>
Table 5.5: Confusion matrices for the visual and audio-visual conditions using network output video. Values represent the percent of responses for each stimulus. Correct responses are in bold.

**Network Video**

<table>
<thead>
<tr>
<th>Stimulus</th>
<th>oh</th>
<th>one</th>
<th>two</th>
<th>three</th>
<th>four</th>
<th>five</th>
<th>six</th>
<th>seven</th>
</tr>
</thead>
<tbody>
<tr>
<td>oh</td>
<td>54</td>
<td>3</td>
<td>16</td>
<td>14</td>
<td>11</td>
<td>0</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>one</td>
<td>0</td>
<td>57</td>
<td>5</td>
<td>22</td>
<td>3</td>
<td>0</td>
<td>14</td>
<td></td>
</tr>
<tr>
<td>two</td>
<td>23</td>
<td>8</td>
<td>49</td>
<td>2</td>
<td>13</td>
<td>4</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>three</td>
<td>0</td>
<td>13</td>
<td>0</td>
<td>79</td>
<td>6</td>
<td>0</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>four</td>
<td>2</td>
<td>4</td>
<td>2</td>
<td>13</td>
<td>70</td>
<td>9</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>five</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>20</td>
<td>7</td>
<td>37</td>
<td>0</td>
<td>33</td>
</tr>
<tr>
<td>six</td>
<td>10</td>
<td>3</td>
<td>5</td>
<td>21</td>
<td>5</td>
<td>3</td>
<td>31</td>
<td>23</td>
</tr>
<tr>
<td>seven</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>8</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>90</td>
</tr>
</tbody>
</table>

**Degraded Audio with Network Video**

<table>
<thead>
<tr>
<th>Stimulus</th>
<th>oh</th>
<th>one</th>
<th>two</th>
<th>three</th>
<th>four</th>
<th>five</th>
<th>six</th>
<th>seven</th>
</tr>
</thead>
<tbody>
<tr>
<td>oh</td>
<td>49</td>
<td>0</td>
<td>36</td>
<td>4</td>
<td>9</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>one</td>
<td>0</td>
<td>77</td>
<td>0</td>
<td>16</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>two</td>
<td>0</td>
<td>0</td>
<td>90</td>
<td>0</td>
<td>10</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>three</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>98</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>four</td>
<td>0</td>
<td>0</td>
<td>6</td>
<td>23</td>
<td>70</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>five</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>37</td>
<td>2</td>
<td>53</td>
<td>0</td>
<td>7</td>
</tr>
<tr>
<td>six</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>100</td>
<td></td>
</tr>
<tr>
<td>seven</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>100</td>
</tr>
</tbody>
</table>

erroneous tongue protrusion frames occurred when there was a significant high frequency component in the audio signal, where the energy in fricatives is concentrated. This type of error might therefore be reduced by allocating more network resources (more input units representing finer acoustic bins) to the high-frequency range, which in the current model was very sparsely represented. Humans are able to easily distinguish such sounds even when provided only with frequencies up to about 3000 Hz (e.g., on the telephone), however. A more human-like solution might therefore be to encourage the recurrent
network to pay more attention to temporal cues, for example, by explicitly requiring the network to estimate the articulatory state some number of frames in the past and future. This could help make the network better able to distinguish the high-frequency sounds on the basis of their surrounding context (Elman, personal communication).

The second systematic difference between the network and original video responses was an excessive number of false “seven” responses to the network video (especially for the stimuli “one,” “five,” and “six”). These responses may have been elicited by the occasional lack of smoothness in the network video. Whenever successive frames were obviously distinct, as might happen with prediction errors, the mouth appeared to make a very sudden movement. This may have been mistaken for a “seven” as that digit was the only one with two syllables and correspondingly fast lip movements. This type of error might be reduced by imposing a smoothness constraint on the network output images, or, again, by encouraging the network to find its own smoothness constraints by forcing a greater sensitivity to temporal cues in the manner suggested above. It is worth noting that the number of false “seven” responses was significantly reduced in the presence of the audio cue, however, even though there were many false “seven” responses to the audio alone. Apparently, while either cue alone was ambiguous, the combination of the two was less so.

In summary, the network output improved comprehension across all stimuli in much the same manner as did the original video. But a general lower level of quality exacerbated the lipreading difficulties present in the original video, and two types of common errors — too frequent estimation of tongue protrusion and occasional loss of smoothness — caused some systematic misinterpretations.
Although the mapping from speech acoustics to vocal tract shape is theoretically non-unique, the studies described here found a high degree of consistency in the mapping, in agreement with other studies of human speech data (e.g., Ladefoged et al., 1978; Papçun et al., 1992; Hogden et al., 1996). This consistency allowed many aspects of articulation to be estimated from acoustics. In those cases in which the static mapping was not consistent, that is when an acoustic window did not unambiguously identify the corresponding articulatory state (as indicated by the failure of a feed-forward network to solve the mapping), it was found that the incorporation of dynamic information by means of recurrent network connections resolved many mapping ambiguities. For a simple recurrent network (SRN) trained on a corpus of voiced stops in a schwa context, average correlations between the most critical estimated and target principal component load trajectories were 0.88 and average root mean squared error (RMSE) was 0.11 (or about 14% of the range). When the principal component loads were converted into movies, the articulators clearly displayed the appropriate gestures: bilabial closure for the /b/, tongue back raising for the /g/, and tongue front raising for /d/. When trained on the more complex “numbers” corpus the SRN achieved average correlations of 0.78 and an average RMSE of 0.09 (11% of the range).

Analysis of errors on two corpora of voiced stops showed that the networks were unable to accurately estimate variability in the position of non-critical articulators, a phenomenon also found by Papçun et al. (1992). Errors of this type suggest that there may be fundamental limits on the extent to which the position of certain articulators can be accurately estimated from acoustics, and therefore that acoustic measurements may
never replace direct measurement for complete representation of the articulators. However, the inherent tendency to average over mappings that are ambiguous caused the networks to estimate plausible, if not accurate, articulatory configurations for non-critical articulators. For example, in the case of an unusually small lip aperture in the production of /ægæ/ the networks estimated a wider aperture more typical for that utterance. In this sense the network can be said to have estimated a canonical rather than actual lip shape, which may be sufficient for many inverse problem applications.

For the “numbers” corpus, the quality of the movies generated from the lip principal component loads was assessed by presenting them to human subjects for speechreading. Digits were correctly identified 59.4% of the time, a level much greater than chance (12.5%) and representing 70% of the intelligibility level afforded by viewing the original lip video. In a preliminary assessment of whether such movies might be used to improve the intelligibility of speech for persons with hearing loss, the audio signal from the spoken digits was low-pass filtered such that correct identification from the degraded audio alone was below 50%. Under these conditions of simulated hearing impairment, digit recognition scores increased to over 80% correct with the network output video (or 86% of the level achieved with original video), indicating that the system holds some promise as an assistive listening device — at least in situations in which a clean audio signal can be channeled to a computer and in which the face of the speaker is not otherwise reliably visible, for example, on the telephone or in a classroom. Analysis of confusion matrices indicated that the greatest number of errors were made on sounds with significant high-frequency components such as fricatives.

It must be recognized that these findings are limited in that the neural networks were trained on only a single speaker and on speech corpora of limited size and variability (although of greater variability than other empirical inverse problem studies to date). It is expected that attempts to “scale up” the method to speaker independence, an unlimited
vocabulary, variable speaking rates, variable acoustic conditions, etc., will meet with some of the same difficulties faced by automatic speech recognition (ASR) systems. But it should be pointed out that this task is not as difficult as that of ASR, principally because it does not require parsing into discrete units. One of the major challenges in ASR is deciding what the discrete units should be (sub-phonemic features, phonemes, diphones, syllables, etc.) and how to break up the audio signal into segments that map onto those units. The gesture recognition task, on the other hand, maps one continuously varying entity (spectral energy) onto another (articulatory positions), so that no explicit segmentation is required. We are also encouraged by the (admittedly small number of) studies that have compared the acoustic-articulatory relationship across subjects and found that there do appear to be systematic regularities (Ladefoged et al., 1978; Papçun et al., 1996). It should also be noted that in some situations training the system on individual speakers might be acceptable, such as for family members or associates of persons with hearing loss; in only several minutes of speech every phoneme in the language may be expressed in an extremely wide variety of contexts. Nonetheless the difficult issue of scaling up is taken seriously, and the reader is encouraged to recognize the limited nature of the results presented here.

A final finding of this study was independent of the neural network results: the articulatory encoding method based on principal component analysis of raw ultrasound and video images was found to provide a potentially useful mechanism for speech production analysis. Ultrasound imaging allows for the non-invasive measurement of tongue surface contours with minimal obstruction of natural articulation, but produces a large number of video images that can be difficult to analyze. The automatic extraction of tongue surface contours by means of edge detection algorithms has proved challenging because the tongue surface line may fade or be obscured by the appearance of edges that do not represent the tongue surface (such as reflections from the palate near regions of
tongue–palate contact). The automatic extraction of principal components and the representation of the images in terms of loadings on these components produces a format in which articulatory trajectories can be viewed and analyzed. In the studies presented here, variability due to coarticulation and other subtle speech production variabilities and asymmetries were found. Although variation in the principal component loads may be non-linearly related to variations in physical space, it is suggested that principal component representations could be converted into Cartesian representations by manually measuring a few strategic images and estimating the tongue contours of other images by non-linear interpolation. Future studies are planned to assess the accuracy of this measurement technique.

**Summary of Experimental Findings**

In summary, the major experimental findings of this (speaker-dependent, limited vocabulary) study may be stated as follows:

- articulatory parameters were estimated from acoustics to within about 11–14% of their range of variability;
- recurrent connections allowed the network to estimate articulatory gestures in cases where the static mapping from acoustics was ambiguous;
- canonical estimations of idiosyncratically varying non-critical articulators reduced estimation accuracy but may not reduce usefulness in certain applications;
- the articulatory estimation and display method may be useful as a special situations assistive listening device for persons with hearing loss; and
- principal component analysis of raw ultrasound and video images may provide a convenient and insightful method for speech production analysis.
Implications for Theories of Human Speech Perception

The accumulating evidence for the feasibility of estimating articulation from acoustics makes ever more pressing the question of whether humans perform some sort of articulatory estimation in the process of speech comprehension. In a hotly contested debate, supporters of the Motor Theory (Liberman et al., 1967; Liberman & Mattingly, 1985) and of Direct Realism (Fowler 1986, 1996) have argued that listeners do infer (or directly perceive) articulation, and supporters of acoustic theories (e.g., Diehl & Kluender, 1989; Lindblom, 1996; Ohala, 1996) have argued that listeners do not. On the motor side of the debate the central argument has been that “perception mirrors articulation more closely than sound” (Liberman, 1967, p. 453). The classical examples in support of this argument are /di/ and /du/ on the one hand, which contain a shared auditory percept (/d/) and shared articulatory gesture (raising of the tongue tip) but very different acoustic signals (a rising vs. falling transition in the second formant), and /pa/ and /ki/ on the other hand, which contain distinct auditory percepts (/p/ vs. /k/) and distinct gestures (labial vs. velar stop), but similar acoustic signals (a burst at 1440 Hz before the vowel). Opponents of motor theories have countered by arguing 1) that an inability to speak does not prohibit language comprehension in infants and mutes (Lenneberg, 1962; MacNeilage, Rootes & Chase, 1967; Ohala, 1996), 2) that humans can discriminate many sounds for which they have no hope of knowing the causal origin, such as bird song and machine sounds (Ohala, 1996) and conversely that animals can distinguish human speech sounds (Kluender, Diehl & Killeen, 1987), and 3) that if the auditory system garners enough information to permit gestural recovery then by an argument from parsimony we should posit that it performs speech discrimination directly, without help from articulation (Lindblom, 1996). Motor theorists have rebutted in turn that 1) the link between speech perception and speech production must be innate, 2) that experimental evidence shows many distinctions between the perception of speech sounds
(processed by the innate module) and non-speech sounds (processed elsewhere), and 3) that the argument from parsimony actually weighs in on the side of the motor theory, for why have two distinct systems that perform such integrated functions (Liberman & Mattingly, 1985). ¹

A New Hypothesis

In the neural network models of the current study, the relationship between acoustics and articulation is not “innate” (i.e., available without experience) but learned through the repeated experience of simultaneously occurring acoustic and articulatory activity. If taken seriously from a psychobiological point of view, the models suggest that articulatory gestures are inferred from acoustics (in agreement with motor theory) but that the acoustic-articulatory relationship is entirely learned through experience (contradicting that theory). The inconsistency of this position vis-à-vis the traditional motor–acoustic debate has led to a new empirically testable hypothesis about the role of articulatory representations in speech perception.

As a preamble to the hypothesis let us establish that speech perception is influenced by non-auditory as well as auditory stimuli. This is demonstrated to some degree by the fact that visual access to the face of a speaker improves speech

¹There have been a number of other arguments put forth on either side of the motor-acoustic debate. See, for example, Lane, 1965; Liberman et al., 1967; Massaro, 1974; Schouten, 1980; Liberman & Mattingly, 1985; Browman & Goldstein, 1986; Diehl & Kluender, 1989; Liberman, 1996; Lindblom, 1996. For collections, see Liberman, 1996; the series of articles in the Journal of Phonetics, 14, in response to the target article by Fowler, 1986; and the series of articles in the special report of the Journal of the Acoustical Society of America, 99 (3), on Speech Recognition and Perception from an Articulatory Point of View, introduced by McGowan and Faber, 1996.
comprehension in noisy conditions (Summerfield, 1979; also, Chapter 5), and also by the fact that synthetically altered visual input can affect the immediate auditory percept. For example, hearing /ba/ while viewing the mouth of a speaker saying /ga/ causes one to perceive /da/ (McGurk & MacDonald, 1976). A biological mechanism for activity in one modality to influence activity in another is Hebb’s (1949) correlational learning, in which synaptic efficacy is regulated by the correlation between pre- and postsynaptic activities. (“Neurons that fire together wire together.”) On this view, the McGurk effect is thought to occur because certain auditory and visual events have been highly correlated throughout one's experience. Specifically, auditory representations of spectral energy fluctuations have become associated with visual representations of lip movements, much as in the artificial neural network studies presented above.

If speech-correlated visual activity can influence speech perception, it is only a small step to argue that other speech-correlated events might similarly become associated with speech. Other candidates include the processes of speech production, because we hear ourselves when we talk. These processes include articulatory motor planning, motor execution, proprioception, and somatosensation. Let us refer to the neural activity underlying such speech-correlated events as speech-correlated neural activity. With this notion defined, we can now formally state the hypothesis:

*Non-auditory (e.g., visual, motor, proprioceptive) speech-correlated neural activity plays an important role in speech perception even when the stimulus is solely acoustic.*

This hypothesis shares with the Motor Theory the notion that motor representations are important in speech comprehension, but differs from that theory in two important respects. First, motor representations are not argued to be primary, or even
necessary, for speech comprehension. Rather, all neural activity that is temporally correlated with speech (more precisely, with the auditory neural activity induced by speech) participates in speech perception, as long as the correlation exists on a regular basis and over a period of time sufficient for the Hebbian mechanism to act. It is therefore acknowledged that the type of activity involved in speech perception may differ by individual, as well as over the time course of a single individual’s development. For example, mute persons and infants must rely more heavily on acoustic and visual activity than on motor activity or proprioception. Second, the perceiver is not presumed to have a separate innate module for processing speech. Differences in the perception of speech and non-speech sounds stem instead from the simple fact that speech sounds tend to correlate with more neural activity than non-speech sounds. To put it simply: we hear and see and plan and execute and feel a small set of speech sounds repeatedly throughout our lives, while most non-speech sounds are experienced in a less repetitive, less rich manner. The hypothesis thus represents an alternative to both acoustic and motor theories of speech perception — motor representations are thought to play a role in the speech perception of normal speaking individuals, but not because of an innate “speech module” or because language can only be transmitted in the code of a gestural score, but instead because of basic principles of neural information processing.

The hypothesis can be broken down into two parts, with the validation of each requiring different investigative methods. Part 1 consists of the claim that speech-correlated activity in any modality may influence speech perception given a non-auditory external stimulus (e.g., visually presented moving lips). This has been clearly shown for visual stimuli using the McGurk effect, as discussed above. Testing whether motor or proprioceptive activity influence speech perception is rather more difficult because it must necessarily be tested with one's own speech. Clearly this situation is different from that of trying to perceive the speech of others as one knows what one is saying even
before one says it. We do, nevertheless, hear ourselves speak, and with some ingenuity this part of the hypothesis is testable. The desired test is that of a “proprio-motor McGurk effect” in which the sensation of producing one sound is time-aligned with the sensation of hearing another.

Houde & Jordan (1998) found such an effect (although they did not name it as such) in an experiment designed to test sensorimotor adaptation to altered acoustic feedback. In this experiment the first two formants of a spoken vowel were shifted in real time so that (for example) when the subject said “pep” she heard “peep.” It was found that speakers compensated for the acoustic alterations by saying something closer to “pop” (the opposite direction in formant space), which caused the altered feedback to sound more like the target “pep.” Importantly, most subjects compensated for the alterations incompletely and were nevertheless unaware that any alteration is taking place. In other words, they perceived “pep” even though they produced something like “pap” (intermediate in formant space between “pep” and “pop”) and the acoustic stimulus was something like “pip” (intermediate between “peep” and “pep”). Just as in the original McGurk effect, the subjects received conflicting information in different modalities — instead of auditory and visual, in this case auditory and proprioceptive. And just as in the original effect, the resulting percept was intermediate between (i.e., largely compatible with) the two types of stimuli. (In this case the percept was intermediate in F1-F2 formant space.) Thus, just as the original McGurk effect illustrated an influence of vision on speech perception, this experiment illustrated an influence of articulatory proprioception (and possibly other speech production related neural activity) on speech perception. (It is possible, however, that the altered percept was due to expectations of hearing a particular sound rather than production related neural activity. An experiment to control for this confounding factor is suggested below.)
Part 2 of the hypothesis claims that non-auditory (e.g. visual, motor, proprioceptive) speech-correlated activity plays an important role in speech perception even when the stimulus is solely acoustic. In other words, it is suggested that the associations developed between auditory and non-auditory neural activity become so strong that acoustic stimulation alone may activate non-auditory neurons (i.e., neurons that previously subserved non-auditory functions). For example, auditory neurons excited by the acoustic stimulus /də/ may in turn excite some motor area neurons typically involved in tongue raising, some proprioceptive neurons typically excited by a raised tongue and slightly open mouth, and some somatosensory neurons typically associated with feeling the roof of the mouth against the tongue. Such activations are not so great in number or magnitude that they lead to tongue raising or cause a conscious sensation of a raised tongue. And they are not expected to occur in individuals with limited experience of the production of the /də/ syllable. In competent speakers, however, they are hypothesized to occur due to the long term effect of the Hebbian mechanism acting on highly correlated neural firings.

Once activated, it is argued that these non-auditory neurons may send information back to the auditory areas or to other areas associated with speech perception. Neural activity flowing through non-auditory areas is not viewed as a passive transmission of the information encoded therein, however. Instead, each modality is thought to offer its own set of constraints so that the information content of the neural signals is changed in the activation flux. The basic idea here is that of interactive activation across modalities leading to an interpretation that is consistent with all modalities (McClelland & Rumelhart, 1981; McClelland, Rumelhart & Hinton, 1986). For example, one might imagine a noisy acoustic stimulus that activates auditory neurons consistent with multiple phonemic interpretations. When this activation spreads to proprioceptive areas (for instance), an ambiguous (not typically encountered) ensemble of proprioceptive neurons
may be excited, which the proprioceptive area may reject as unstable (that is, mutual excitation and inhibition within the proprioceptive area may cause the firing patterns set in motion by the auditory activity to be changed). The altered proprioceptive activity may then feed back to the auditory area which may then favor one phonemic interpretation over another. In this way, the somatosensory activity may “clean up” the auditory activity, and thereby improve the capacity for accurate speech perception.

Supporters of acoustic theories of speech perception argue that the “cleaning up” that has been described here as taking place in somatosensory areas could just as well happen directly in auditory areas. The proposed hypothesis does not deny this possibility. In fact the hypothesis suggests exactly this for listeners who have limited experience with speech production. But for accomplished speaker-listeners, a growing body of experimental evidence indicates an important role for non-auditory areas in the process of speech comprehension.

Before reviewing this evidence, let us first address the issue of parsimony — the argument that articulatory representations should not be posited to play a role in speech perception unless they are absolutely required. In Lindblom's (1996, p. 1690) words: “... 'parsing' the signal into its articulatory components seems unnecessary because, if it contains enough distinctive (auditory) information to permit recovery of gestures, it also ought to contain enough (auditory) information to permit lexical access without going by way of articulation. Why make a detour ...?” (parentheses in original). The important issue here is that the hypothesized correlational learning is not at all effortful or planned at some high organizational level. It is an entirely local phenomenon caused by neurophysiological processes. Indeed it is not clear that even at the level of the organism the hypothesized correlations could be willfully ignored. It should therefore not be considered a “detour” or “extra effort” to take note of these correlations. It is not clear that an argument from parsimony applies at all to the human brain, which is an evolved
rather than a designed structure. Indeed, redundancy seems to be at least as important an organizational principal of the brain as is parsimony.

**Behavioral and Neuropsychological Evidence for the Hypothesis**

The evidence in support of the hypothesis is multifaceted. First, the McGurk effect discussed above shows unambiguously that there is a tight interaction between audition and at least one other modality, namely vision. It is important to underline here that the basic auditory percept is altered in this effect such that the listener unequivocally perceives a different sound in the presence of the altered visual stimulus. Second, cortical stimulation studies in epileptic surgical patients have shown a disruption of sequential orofacial movements and phoneme identification from the same brain sites (Ojemann & Mateer, 1979). Third, activation of Broca’s area, which is classically associated with speech production, has been found in an auditory phonetic discrimination task (Zatorre, Meyer, Gjedde & Evans, 1996). Conversely, activation of auditory cortex has been found during silent lipreading (Calvert et al., 1997). Fourth, if speech production areas were not involved in speech perception, we might expect cases of pure production aphasia to be relatively common. Such cases rarely if ever occur, however, especially in the acute phase (Berndt, personal communication; Mohr et al., 1978). (Pure production aphasia in the chronic phase leaves open the possibility of reorganization of speech perception processes so that they are less dependent on production-related activity, as is proposed for mutes.) The rarity of pure production aphasia indicates that damage to production areas tends to disrupt speech perception as well. Fifth, analogous cross-modal interactions have been found in non-speech tasks. For example, premotor neuronal activity has been found in response to the visual observation of hand and mouth movements in macaques (Gallese, Fadiga, Fogassi & Rizzolatti, 1996; Rizzolatti, Fadiga, Gallese & Fogassi, 1996), and increased motor evoked potentials have been found in humans during the
visual observation of hand and mouth movements (Fadiga, Fogassi, Pavesi & Rizzolatti, 1995). Finally, the “proprio-motor McGurk effect” (Houde & Jordan, 1998; Moody, 1998) seems to indicate that motor and proprioceptive activity may influence speech perception much as in the visual McGurk effect (although this claim still lacks control of the possible confound of listener expectations, addressed below). Thus, the suggestion from the modeling studies that articulatory extraction is feasible combines with a host of behavioral and neuropsychological findings to suggest not that non-auditory recruitment is a logical necessity, but rather that it is a biological reality.

**Experimental Predictions of the Hypothesis**

The new hypothesis makes a number of falsifiable experimental predictions, including:

- **Neuroimaging studies of normal speaking vs. speech-disabled persons** (e.g., early muted individuals) in speech listening tasks will show significant differences in the two populations due to the reduction of the participation of speech production activity in the speech-disabled.

- **Motor evoked potentials (MEPs) in speech production musculature** will differ under the conditions of listening to speech vs. non-speech stimuli (analogous to MEP studies of the hand performed by Fadiga et al., 1995). MEPs will also differ during speech and non-speech visual tasks such as viewing speechreadable facial movements vs. viewing hand motions not related to speech.

- **Single-cell recordings in the vocal production control areas of the brains of vocal animals** (e.g., non-human primates, songbirds) will show activity in response to acoustic stimuli.
The Houde paradigm for altered acoustic feedback will illustrate a “proprio-motor McGurk effect” in an echo condition designed to control for listener expectations. Recall that the phenomenon of interest in this task is that a subject hears “pip” but perceives “pep.” The two competing explanations for this are that the subject perceives “pep” (1) because she expects to hear “pep” or (2) because her acoustic feedback indicates “pip” and her proprioceptive feedback indicates “pap” and the integration of these two types of feedback settles on an intermediate (in formant space) percept “pep.” A possible control is simply to play the acoustic feedback of each utterance to the subject twice, once coincident with the production of the utterance and once a second or so later. The subject's expectations should be the same during both utterances (as the subject will have learned to expect a repetition of her speech during the baseline phase of the experiment in which the acoustics are unaltered). Speech production related neural activity will be different in the two phases, however, and so the hypothesis predicts that the subject will perceive the two repetitions differently.

In short, the proposed hypothesis is rich in the number and variety of experimental predictions that it makes. The hypothesis is also timely in that it illustrates a larger trend in the cognitive sciences: the tendency for knowledge at the neural level to directly impact theories at the behavioral level. If this hypothesis (which is framed at the neural level) can be verified, it will help explain (at the behavioral level) non-auditory influences on speech perception.
REFERENCES


