Kinematic Measurement and Feature Sets for Automatic Speech Recognition

Thesis by

Daniel C. Fain

In Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy

California Institute of Technology

Pasadena, California

2001

(Defended May 14, 2001)

Acknowledgements

My dear and faithful friend, Dr. Amy L. Greenwood, was instrumental in helping me complete this dissertation, keeping me going, discussing research, and reading drafts. I cannot thank her enough. Members of my family who have inspired me by academic example are my late, beloved, and multitalented mother; my dear sister, whose dissertation research involves snorkeling on Maui to adjust her instruments; my uncles, father, and late grandfather and great-grandfather.

Several other of my close friends supported me both personally and by discussing the basis of the work described here: the brilliant Dr. Michael S. Wehr, reverse engineer of the brain; the unstoppable Mr. Adam R. Woodbury, deriver of equations of motion of grass in the wind, and of fog; the imperturbable Dr. Patrick Y.-S. Chuang, master of the math of the forming clouds; the indefatigable Dr. Shaun S. Shariff, modeler of fish and flame; and the incomparable Ms. Bena L. Currin (how I wish I knew even half what Bena does about light). Adam also contributed images for my candidacy presentation.

My advisor, Prof. Alan H. Barr, suggested working on the inverse-problem approach to speech recognition, and provided guidance and support along the way. I thank Prof. Mark Konishi for bringing me into his neuroethology research group for a year and serving on my candidacy committee. Also generously making time for my candidacy meeting was Prof. Carver Mead. Doctors Doug Kerns and Ron Benson brought me to Tanner Research, Inc., where the work described in chapter 2 was performed. As principal investigator of that project, I enlisted the help of Dr. Garry Chinn at an early stage; his role is described further in section 1.13. Professor Andreas Andreou kindly included me in his acoustic processing group for a 1997 workshop at Johns Hopkins (appendix B). Professor John Hopfield and his students and postdocs welcomed me to their very edifying group meetings. I had extremely valuable discussions with my friend Dr. Sam T. Roweis, and he also provided corrected microbeam data for the experiments of chapter 3. Professor Joe Picone of Mississippi State University and Dr. Aravind Ganapathiraju of Conversay Inc. responded immediately to my request

for data used in chapter 4. My friend Prof. Erik Winfree helped straighten out the ideas of chapter 4; and Ms. Maria Silke, the statistics in chapter 3. I thank my thesis committee, Professors Alan H. Barr, Yaser Abu-Mostafa, Abeer Alwan (UCLA), James Arvo, and Pietro Perona, the first three of whom were also on my candidacy committee.

Thanks also to the people of the Computation and Neural Systems program including Dr. Kurt Fleischer and Dr. David Rosenblum; to members of the Computer Graphics Group not previously mentioned including Dr. David Laidlaw, Mark Montague, Louise Foucher, and Dave Felt; to the researchers in Tanner Labs including Dr. Mass A. Sivilotti, Dr. John Tanner, Dr. Tom Bartolac, and Lee Fisher; and to Dr. Tim Anderson of the Air Force Research Laboratory for acting as our technical point of contact for the work described in chapter 2. Various people in the Graphics Group, especially Dr. Ronen Barzel, created the LaTeX code used to format this thesis and generate the iconic index of figures.

The author was supported by a National Science Foundation (NSF) Graduate Research Fellowship; a National Institute of Mental Health Training Grant; SBIR Contract F41624-97-C-6017; the NSF and Defense Advanced Research Projects Agency (DARPA) Science and Technology Center for Computer Graphics and Scientific Visualization; Johns Hopkins University; and by equipment grants from the Office of Naval Research, Hewlett-Packard, and Intel.

Abstract

This thesis examines the use of measured and inferred kinematic information in automatic speech recognition and lipreading, and investigates the relative information content and recognition performance of vowels and consonants. The kinematic information describes the motions of the organs of speech—the articulators. The contributions of this thesis include a new device and set of algorithms for lipreading (their design, construction, implementation, and testing); incorporation of direct articulator-position measurements into a speech recognizer; and reevaluation of some assumptions regarding vowels and consonants.

The motivation for including articulatory information is to improve modeling of coarticulation and reconcile multiple input modalities for lipreading. Coarticulation, a ubiquitous phenomenon, is the process by which speech sounds are modified by preceding and following sounds.

To be useful in practice, a recognizer will have to infer articulatory information from sound, video, or both. Previous work made progress towards recovery of articulation from sound. The present project assumes that such recovery is possible; it examines the advantage of joint acoustic-articulatory representations over acoustic-only. Also reported is an approach to recovery from video in which camera placement (side view, head-mounted) and lighting are chosen to robustly obtain lip-motion information.

Joint acoustic-articulatory recognition experiments were performed using the University of Wisconsin X-ray Microbeam Speech Production Database. Speaker-dependent monophone recognizers, based on hidden Markov models, were tested on paragraphs each lasting about 20 seconds. Results were evaluated at the phone level and tabulated by several classes (vowel, stop, and fricative). Measured articulator coordinates were transformed by principal components analysis, and velocity and acceleration were appended. Concatenating the transformed articulatory information to a standard acoustic (cepstral) representation reduced the error rate by 7.4%, demonstrating across-speaker statistical significance ($p = 0.018$). Articulation improved recognition of male speakers more than female, and recognition of vowels more than fricatives or stops.

The analysis of vowels, stops, and fricatives included both the articulatory recognizer of chapter 3 and other recognizers for comparison. The information content of the different classes was also estimated. Previous assumptions about recognition performance are false, and findings of information content require consonants to be defined to include vowel-like sounds.

Contents

Acknowledgements *iii*

Abstract *v*

Index of Figures *xiv*

List of Tables *xxiii*

1 Introduction **1**

1.1 Why Use Articulation in Recognizers? 1

1.1.1 Acoustic Cues 2

1.1.2 Describing Coarticulation in the Articulatory Domain 3

1.1.3 Motor-Space Reconciliation of Lipreading and Hearing Speech 4

1.1.4 Sources of Kinematic Representations used in this Thesis 4

1.2 Articulatory Theories of Human Speech Perception 5

1.2.1 Gesture Theory of Speech 5

1.2.2 Motor Theory of Human Speech Perception 6

1.2.3 Motor Phonetics 6

1.2.4 Articulatory Phonology 6

1.3 Evidence For and Against the Motor Theory 7

1.3.1 Lipreading with Normal Hearing 7

1.3.2 Likely Mechanisms for Human Lipreading Ability 8

1.3.3 Infant Language Development and Motor Representations 8

1.3.4 Magnetoencephalography (MEG) Studies 9

1.3.5 Language Instruction 10

1.4 Articulatory Recognition Proposed 10
1.5 State-of-the-Art Automatic Speech Recognition 10
1.6 Measuring Error Rates 11
1.6.1 Error Rates as the Only Metric: Shortcomings and Advantages 12
1.6.2 Task Specificity 12
1.7 Relation between Lipreading and Articulatory Recognition 13
1.7.1 Approaches to Multimodal Input Integration 13
1.8 Timescales of a Recorded Word 16
1.9 Linguistic Units of Speech 16
1.9.1 Phonemes 18
1.9.2 Allophones 18
1.9.3 Distinctive Features 19
1.9.4 Gestures 20
1.9.5 Syllables 21
1.9.6 Words 21
1.9.7 Importance of Pitch 21
1.9.8 Sentences 21
1.10 Front-End Algorithms in Conventional State-of-the-Art Recognizers 22
1.10.1 Separation into Front and Back Ends 22
1.10.2 Acoustic Analysis Using the Cepstrum 22
1.11 Conventional Back-End Algorithms 26
1.11.1 Markov Models 26
1.11.2 Hidden Markov Models 27
1.11.3 Engineering Units of Speech: Monophones and Triphones 32
1.11.4 Grammar Modeling 32
1.12 Historical Notes 33
1.12.1 Origins of Automatic Speech Recognition 33
1.12.2 Articulatory Symbols for Writing 34
1.12.3 History of Speech Synthesis 34
1.13 Third-Party Assistance 37
1.14 Summary of Results of Thesis 37

2 Side-View Lipreading Device and Algorithms 39

2.1 Problem Description . . . 39

2.2 Background . . . 40

2.2.1 Degrees of Freedom of Lip and Jaw Motion . . . 40

2.2.2 Distinctive Phonetic Features Relating to Lips and Jaw . . . 41

2.3 Previous Work . . . 41

2.3.1 History of Machine Lipreading . . . 41

2.3.2 Different Ways to Merge Audio and Video . . . 42

2.3.3 Dynamic Contours (Snakes) . . . 43

2.3.4 Camera Placement . . . 43

2.4 Lipreading Face Mask . . . 43

2.4.1 Design Issues . . . 45

2.4.2 Alternate Designs Considered: Silhouette Imaging . . . 45

2.4.3 Implementation . . . 45

2.5 Front-End Feature Extraction Using Face Mask . . . 46

2.5.1 Noise Removal with Median Filter . . . 49

2.5.2 Thresholding Face from Background . . . 49

2.5.3 State Machine Distinguishing Upper and Lower Lips . . . 50

2.5.4 Finding Corner where Lips Meet . . . 50

2.5.5 Calculating Centroids of Lip Regions . . . 50

2.5.6 Features Generated by Front End . . . 51

2.5.7 Robustness to Mask Placement . . . 51

2.6 Centroid Motion Versus Lip Motion . . . 52

2.6.1 Aperture Effect . . . 53

2.6.2 Comparison to Previous Methods . . . 54

2.7 Simple Back-End Classifier . . . 55

2.7.1 Single-Frame Recognizer . . . 56

2.8 Lipreading Recognition Results . . . 56

2.8.1 Acoustic Recognizer . . . 57

2.8.2 Still Image Recognizer . . . 58

2.8.3 Analysis of Errors by Modality . . . 59

2.9 Side-View Lipreading: Conclusions . . . 59

2.10 Future Work 59

3 Speech Recognition with Direct Articulatory Measurements 60

3.1 Long-Term Goal 61

3.2 Problem Statement 62

3.3 Related Work 62

3.3.1 Human Speech Perception 62

3.3.2 Interpolating Acoustic Models of Coarticulation 62

3.3.3 Recognition with a Categorical Motor Representation: The Articulatory Feature Model 63

3.3.4 Other Recognizers with Categorical Motor Spaces 64

3.3.5 Recovery of Positions from Sound 64

3.3.6 Articulatory Recovery with Kinematic Models 65

3.3.7 Incorporating Kinetics into Articulatory Recovery 65

3.3.8 Automatic Labeling of Articulatory Events 66

3.3.9 Previous Articulatory Recognition with Measured Data 66

3.4 Direct Kinematic Articulatory Measurements 67

3.4.1 Wisconsin X-Ray Microbeam 67

3.4.2 Alternate Direct Measurement Techniques 68

3.5 Preparation of Data for Recognition 71

3.5.1 Tracking Correction (Previous Work) 71

3.5.2 Removal of Extraneous Sounds 72

3.5.3 Total Duration of Each Speaker's Usable Data 72

3.5.4 Hand Transcription 73

3.5.5 Criteria for Inclusion in Recognition Experiments 74

3.6 Recognizer Training 75

3.7 Front-End Processing of Sound Recording 75

3.8 Front-End Processing of Microbeam Articulatory Data 75

3.8.1 Articulatory Parameterization and Constraints 77

3.8.2 Principal Components Analysis (PCA) 77

3.8.3 First and Second Time Derivatives 78

3.8.4 Normalization 78

3.9 Back-End Recognizer for Acoustics and, Optionally, Articulation 79
3.9.1 Unigram Model for Grammar 79
3.9.2 Training Sequence for Articulatory Recognizer 80
3.9.3 Train/Test Split 81
3.9.4 Flat Start 82
3.9.5 Restarting Training with New Models 83
3.9.6 Optimality of Recognizer Architecture 83
3.10 Recognizer Testing 83
3.11 Setting Global Parameters of Recognition 84
3.11.1 Variance Floor 84
3.11.2 Grammar Weight and Insertion Bias Defined 85
3.11.3 Joint Optimization of Grammar Weight and Insertion Bias 85
3.11.4 Comparison to Other Recognizers 86
3.12 Increased Recognition Performance with Measured Data 87
3.12.1 Recognition of Consonants versus Vowels 87
3.12.2 Number of Parameters Required 88
3.12.3 Variability between Test Paragraphs 88
3.12.4 Performance on Training Data 88
3.13 Summary of Results 90
3.14 Recognition with Microbeam Data: Conclusions and Discussion 90
3.15 Future Work 91
3.15.1 Testing with Inferred Articulation 91
3.15.2 Applying Optimal Feature Extraction 92

4 Recognition and Entropy of Vowels and Consonants 93
4.1 Different Types of Consonants 94
4.2 Entropy of Vowels and Consonants 94
4.2.1 Categorization of Vowels, Stops, and Fricatives 95
4.2.2 Measurement of Entropy 97
4.2.3 First-Order Entropy Estimation 97
4.3 Previous Work on Vowel Versus Consonant Recognition 98
4.3.1 Conventional Recognition Architectures 98

4.3.2 Nonstandard Recognition Architectures 98
4.3.3 Prior Work on Text Analysis 99
4.3.4 Human Listeners 100
4.4 Vowel and Consonant Recognition with Sound and Articulation 100
4.5 Vowel and Consonant Recognition by the Commercially Available ViaVoice Recognizer 101
4.6 Vowel and Consonant Recognition by Recnet Recurrent Neural Network 102
4.7 Vowel and Consonant Recognition on Switchboard Telephone Conversations 102
4.7.1 Determination of Phonetic Errors from Word-Level Transcripts 103
4.7.2 Vowel and Consonant Recognition: Switchboard Results 104
4.7.3 Consequences of the Grammar Model and Dictionary 104
4.8 Recognition and Entropy of Vowels and Consonants: Conclusions 105
4.9 Future Work 105
4.9.1 Calculation of Entropy from Text 105
4.9.2 Textual Experiments with Human Listeners 105
4.9.3 Audio Experiments with Human Listeners 106
4.9.4 Experiments with Other Recognition Architectures 106
4.9.5 Analysis of Recognizer Performance 106

5 Discussion and Conclusions **107**
5.1 Advantages and Disadvantages of the Side-View Approach to Lipreading 108
5.2 Advantages and Disadvantages of Motor Representations in Speech Recognition . . . 108
5.3 How Many Articulatory Degrees of Freedom Are Important? 109
5.4 Discussion: Recognition and Entropy of Vowels and Consonants 110

Appendixes:

A Principal Components Analysis of Microbeam Data **112**
A.1 Definitions: PCA and MLG 112
A.1.1 Principal Components Analysis (PCA) 112
A.1.2 Maximum-Likelihood Gaussian (MLG) Classifier 113
A.1.3 Single Transformation Matrix for All Categories 113
A.1.4 Category-Dependent Transformation Matrices 114
A.2 Previous Work on PCA of Speech Production Data 114

A.2.1 Sagittal Factor Analysis of Tongue (Harshman et al. 1977) 114
A.2.2 Cross-Sectional Tongue Shapes for Vowels (Stone et al. 1997) 114
A.2.3 Across-Class PCA in Wisconsin X-Ray Microbeam Data (Roweis 1999) 115
A.3 Results 115
A.3.1 Rationale for Using Single Transformation Matrix 115
A.3.2 Percent of Variance Explained by Principal Components 116
A.3.3 Statistics for Individual Phonetic Classes 116
A.4 PCA: Conclusions and Discussion 130

B Front-End Optimization **132**
B.1 Optimization of Front-End Linear Processing 132
B.2 Typical Front-End Components 133
B.3 Need for Dimensionality Reduction 133
B.4 Other Dimensionality Reduction Techniques 133
B.4.1 Principal Components Analysis (PCA) 133
B.4.2 Linear Discriminant Analysis (LDA) 134
B.5 Heteroscedastic Discriminant Analysis (HDA) 134
B.5.1 Statistics Required to Perform HDA 134
B.5.2 Baum-Welch Training of HDA 135
B.5.3 HDA and Weighted Sums of Gaussians 135
B.6 Scaling Problems in Speech Recognition 135
B.7 Switchboard Telephone Corpus 136
B.7.1 Large-Vocabulary Continuous Speech Recognition Workshop 136
B.8 Experimental Procedure 136
B.8.1 Generation of Context Windows 137
B.8.2 Segmentation per Triphone Alignment 137
B.8.3 Running HDA 137
B.8.4 Retraining 137
B.8.5 Problems with Variance Floor and Grammar Weight 138
B.8.6 Results 138
B.9 Conclusions 138

C Lipreading Aids for the Hearing Impaired **139**

C.1 Lipreading by the Hearing Impaired 139

D Parameterizations of Lip and Jaw Motion in FACS and MPEG-4 **141**

D.1 Facial Action Coding System (FACS) 141

D.2 Motion Picture Experts Group (MPEG) 4, Synthetic-Natural Hybrid Coding (SNHC) 142

E X-Ray Microbeam Tracking Technology **144**

E.1 Microbeam Tracking as a Substitute for Cineradiography 144

E.2 Tracking Algorithm 145

E.3 Bead Placement 145

F Text Samples **146**

F.1 Paragraphs used for Recognition Experiments 146

F.1.1 "Grandfather" Paragraphs 146

F.1.2 "Hunter" Passage 146

F.2 Example Sentences for Importance of Consonants and Vowels 147

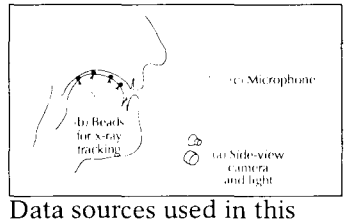
F.3 Random Text Generated by a Markov Model 148

References **149**

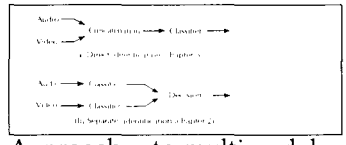
Glossary *166*

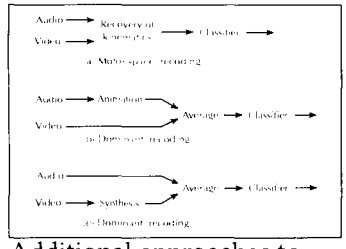
Figures

Chap. 1. Introduction

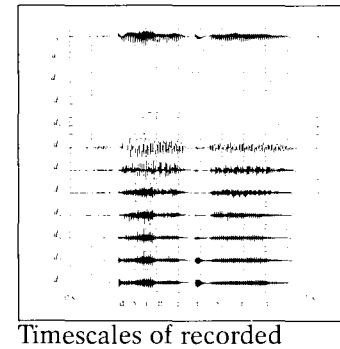
1.1 

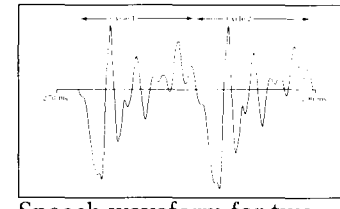
Data sources used in this thesis . . . 5

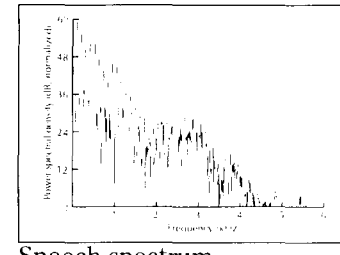
1.2 
Approaches to multimodal integration used in this thesis . . . 13

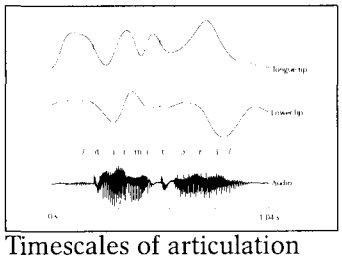
1.3 

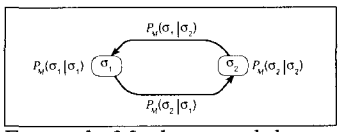
Additional approaches to multimodal integration . . . 14

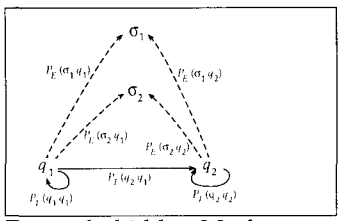
1.4 
Timescales of recorded word: wavelet decomposition . . . 17

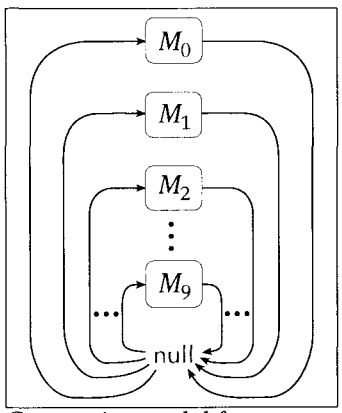
1.5 
Speech waveform for two vocal-cord vibrations . . . 18

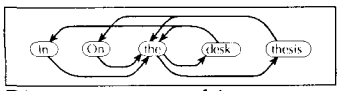
1.6 
Speech spectrum . . . 19

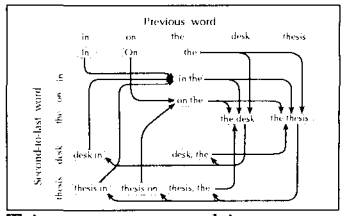
1.7  . . . 20
Timescales of articulation and sound

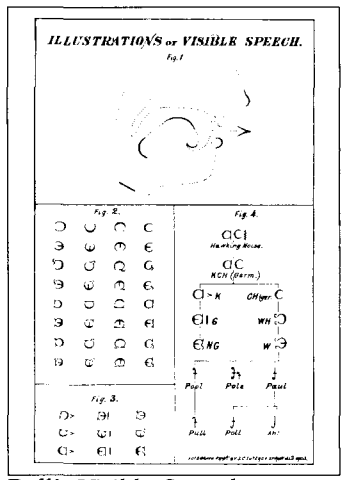
1.8  . . . 26
Example Markov model

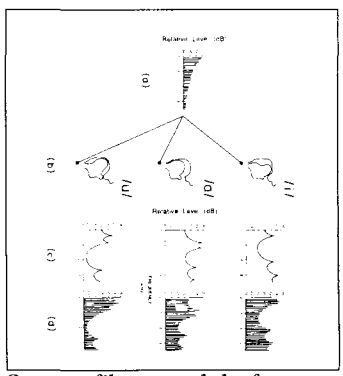
1.9  . . . 29
Example hidden Markov model

1.10  . . . 31
Composite model for continuous digit sequences

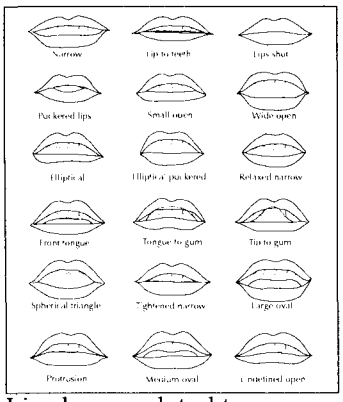
1.11  . . . 33
Bigram state machine

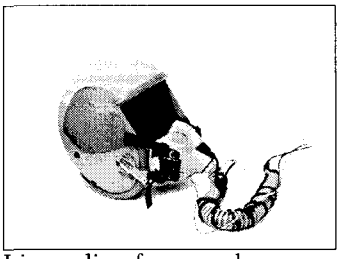
1.12  . . . 33
Trigram state machine

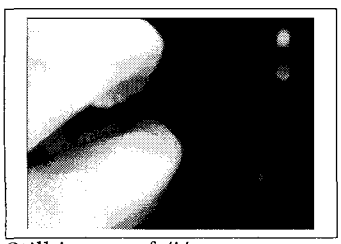
1.13  . . . 35
Bell's Visible Speech alphabet, representing articulations graphically

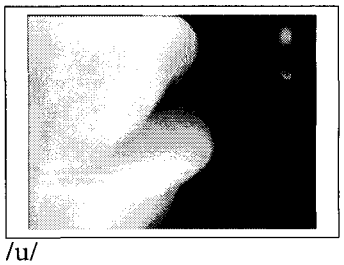
1.14  . . . 36
Source-filter model of speech synthesis. Reprinted with permission (Bailey 1983); copyright 1983, Academic Press

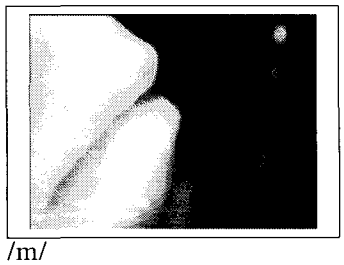
Chap. 2. Side-View Lipreading Device and Algorithms

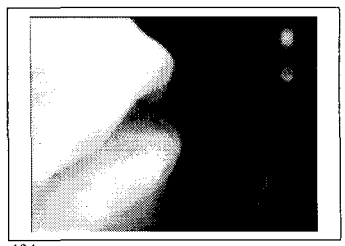
2.1  . . . 42
Lip shapes related to speech. Reprinted with permission (Parke and Waters 1996); copyright 1996, A. K. Peters

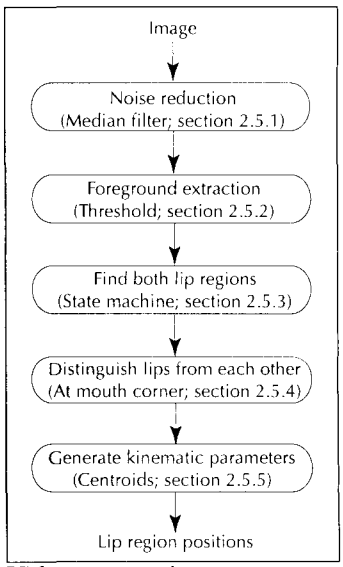
2.2  . . . 44
Lipreading face mask. Copyright Tanner Research; used with permission

2.3  . . . 46
Still image of /i/. Copyright Tanner Research; used with permission

2.4  . . . 47
/u/

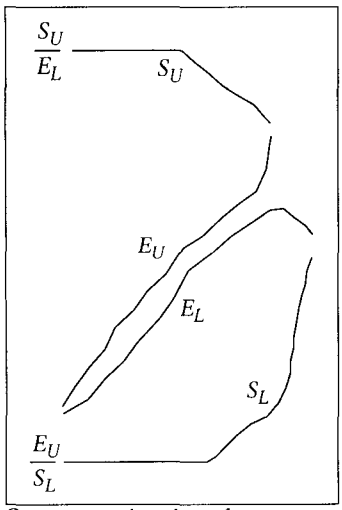
2.5  . . . 47
/m/

2.6  . . . 48
/f/

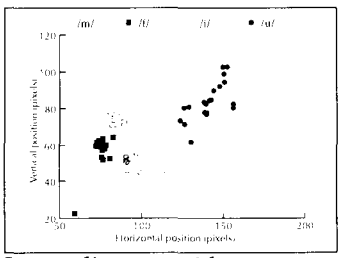
2.7 

. . . 48
Video-processing steps

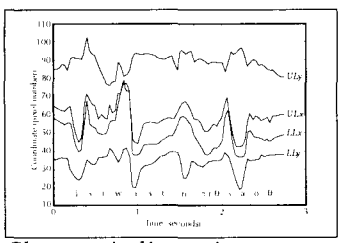
2.8 Pseudocode for lower-lip segmentation. 51

2.9 

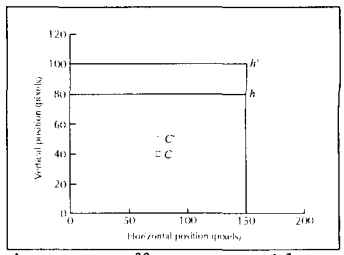
Segmentation into lower, upper lips . . . 52

2.10 

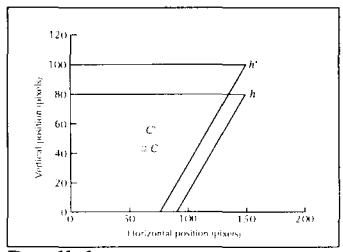
Lower-lip centroid positions for /m/, /f/, /i/, and /u/ . . . 53

2.11 

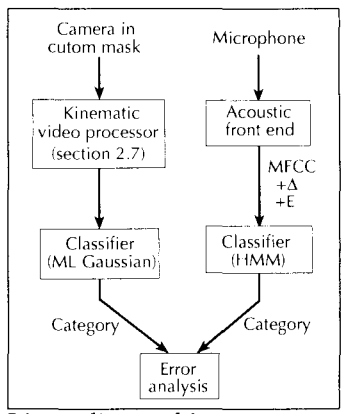
Changes in lip region coordinates over time . . . 54

2.12 

Aperture effect: centroid of shifting rectangle . . . 55

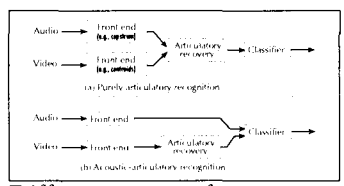
2.13 

Parallelogram aperture effect . . . 56

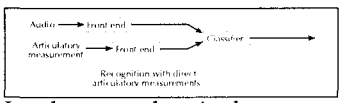
2.14 

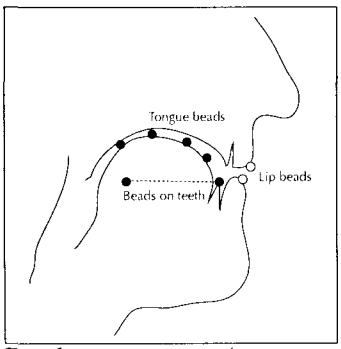
Lipreading architecture . . . 57

Chap. 3. Speech Recognition with Direct Articulatory Measurements

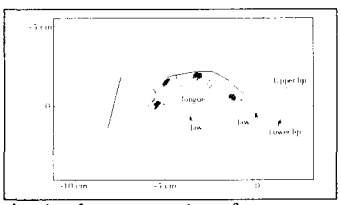
3.1 

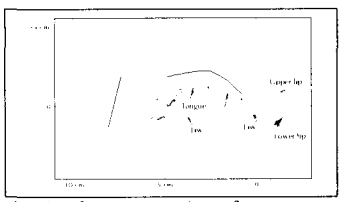
Different types of articulatory recognition . . . 61

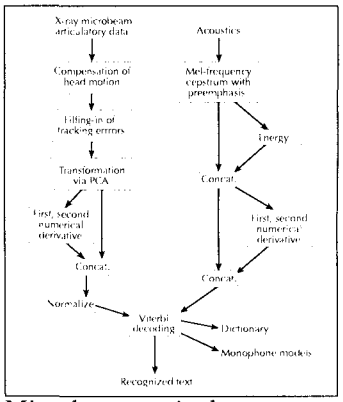
3.2  . . . 62
Implemented articulatory architecture

3.3 

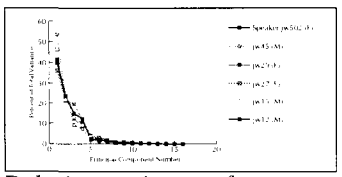
. . . 68
Bead arrangement (same as figure A.1

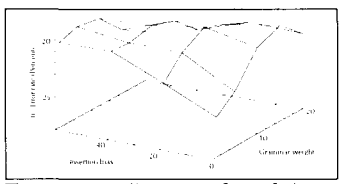
3.4  . . . 69
Articulator motion for "kuh"

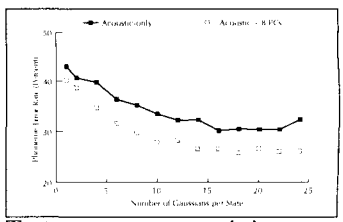
3.5  . . . 70
Articulator motion for "puh"

3.6 

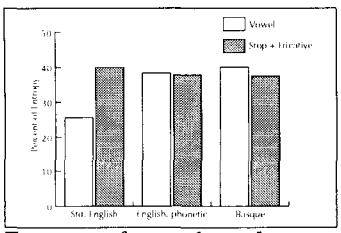
. . . 76
Microbeam articulatory recognition architecture

3.7  . . . 78
Relative variance of principal components

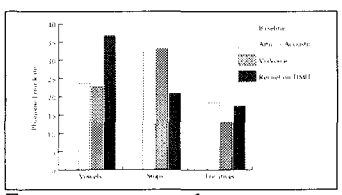
3.8  . . . 86
Error rate (inverted scale) as a function of grammar weight and insertion bias

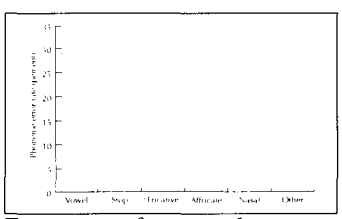
3.9  . . . 88
Test-set error as training progresses and parameters are added

Chap. 4. Recognition and Entropy of Vowels and Consonants

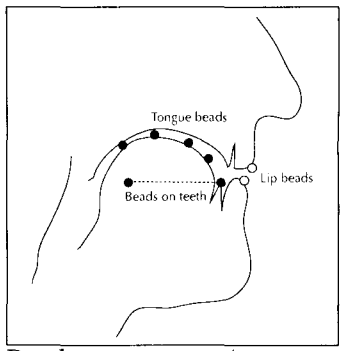
4.1 

. . . 98
Entropy of vowels and consonants in two languages

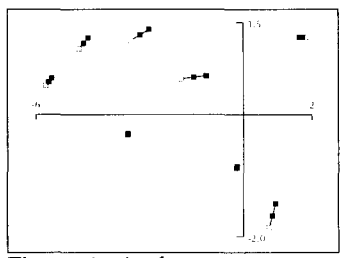
4.2 
. . . 101
Error rate: vowels, stops, fricatives

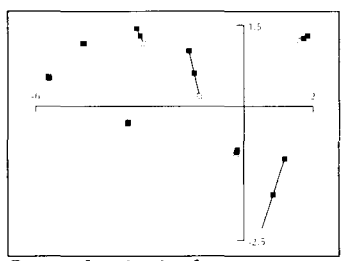
4.3 
. . . 104
Error rates for vowels, stops, and fricatives in automatic recognition of telephone conversations

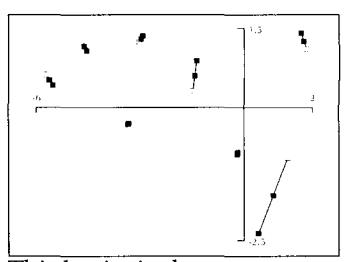
App. A. Principal Components Analysis of Microbeam Data

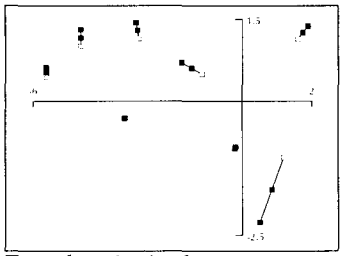
A.1 

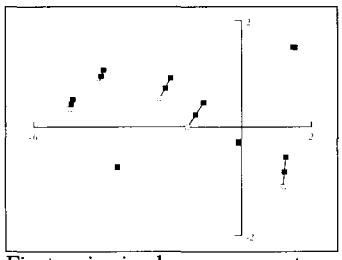
. . . 116
Bead arrangement (same as figure 3.3)

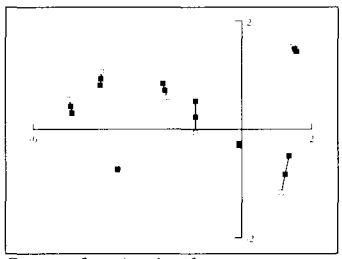
A.2 
. . . 117
First principal component for JW12

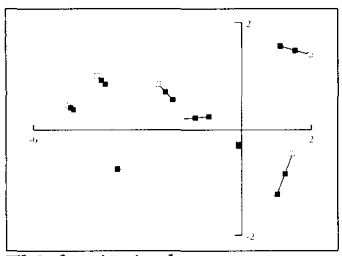
A.3 
. . . 117
Second principal component for JW12

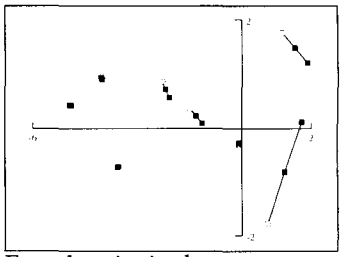
A.4 
. . . 118
Third principal component for JW12

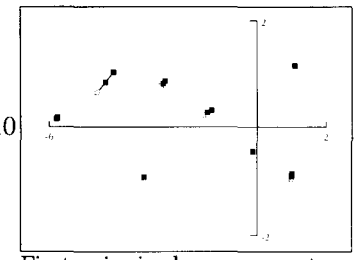
A.5  . . . 118
Fourth principal component for JW12

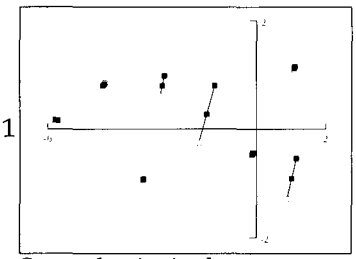
A.6  . . . 119
First principal component for JW15

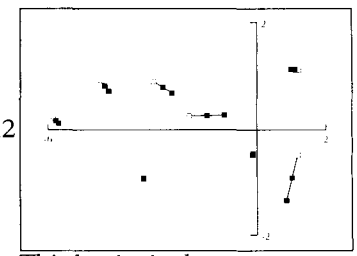
A.7  . . . 119
Second principal component for JW15

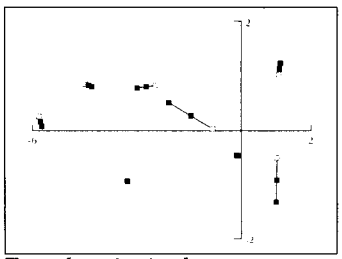
A.8  . . . 120
Third principal component for JW15

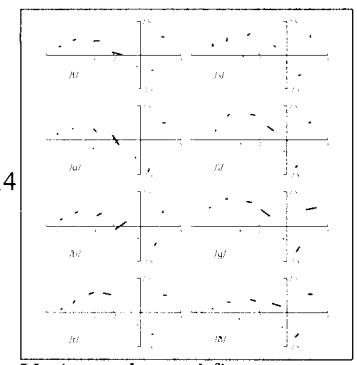
A.9  . . . 120
Fourth principal component for JW15

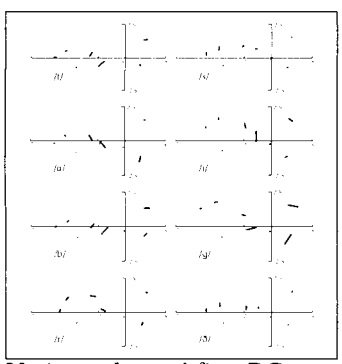
A.10  . . . 121
First principal component for JW27

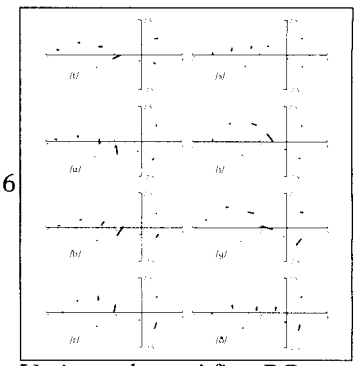
A.11  . . . 121
Second principal component for JW27

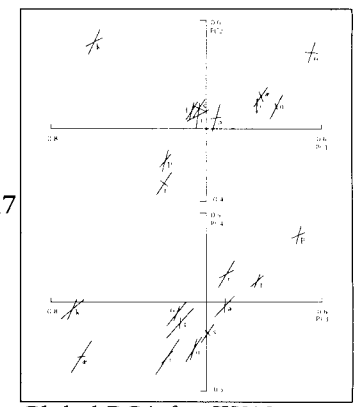
A.12  . . . 122
Third principal component for JW27

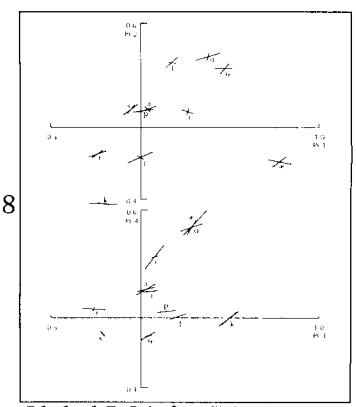
A.13  . . . 122
Fourth principal component for JW27

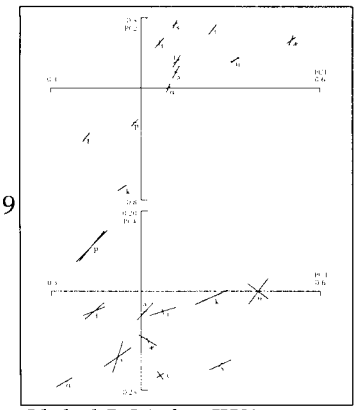
A.14  . . . 124
Various phones' first principal component (PC) for JW12

A.15  . . . 125
Various phones' first PC for JW15

A.16  . . . 126
Various phones' first PC for JW27

A.17  . . . 127
Global PCA for JW12

A.18  . . . 128
Global PCA for JW15

A.19  . . . 129
Global PCA for JW27

Tables

Chap. 2. Side-View Lipreading Device and Algorithms

2.1 Previous work: lipreading error rates by human listeners for different face parameterizations 41
2.2 Confusion matrix for /m/, /f/, /i/, /u/ 58
2.3 Error analysis of audio and video recognition 58

Chap. 3. Speech Recognition with Direct Articulatory Measurements

3.1 Erler and Freeman's discretized articulator states 64
3.2 Per-speaker phone counts 73
3.3 Articulatory recognition subject demographics 74
3.4 Word frequencies for unigram model 80
3.5 Examples of sentences used for training 82
3.6 Articulatory recognition rates by speaker 84
3.7 Cross-validation of grammar weight and insertion bias 87
3.8 Recognition performance with different numbers of parameters 89
3.9 Recognition performance on individual paragraphs 89
3.10 Recognition performance on training data 89

Chap. 4. Recognition and Entropy of Vowels and Consonants

4.1 English letters: vowels, stops, and fricatives 96
4.2 Wisconsin phonetic symbols: vowels, stops, and fricatives 96
4.3 Basque letters: vowels, stops, and fricatives 96

4.4 Vowel and consonant recognition rates: previous work 99
4.5 Error rate reduction (microbeam data): vowels, stops, and fricatives 100
4.6 Recnet error rates: vowels, stops, and fricatives 102
4.7 Switchboard phonetic symbols: vowels, stops, fricatives, affricates, and nasals 103

App. A. Principal Components Analysis of Microbeam Data

A.1 Variance explained by principal components, for all speakers combined 123
A.2 Phonetic contrasts in figure A.14 through figure A.16 130

App. B. Front-End Optimization

B.1 Improved performance on Switchboard with HDA 138

App. D. Parameterizations of Lip and Jaw Motion in FACS and MPEG-4

D.1 Lip and jaw motions in facial action coding system 143

Chapter 1

Introduction

This thesis reports the merit of augmenting sound with kinematic information for automatic speech recognition. The rationale for using such a representation is described in section 1.1; the question whether ideas about motor representations are used in human perception is discussed in section 1.2. The kinematic inputs describe the motion of points on the articulators (organs of speech). In one project, lip motion was recovered from video (chap. 2). In a second project (chap. 3), an existing data set of articulator motions and synchronized sound recordings was used as recognizer input. To evaluate how articulation data affects articulatory recognizers, separate error rates have been calculated in this project for vowel and consonants, using both the articulatory recognizer and some conventional recognizers (chap. 4). The information content of vowels and consonants in text has also been calculated.

1.1 Why Use Articulation in Recognizers?

This thesis uses an inverse-problem, motor representation approach to recognition. Proponents of the motor theory cite the following arguments and evidence (Liberman and Mattingly 1985). If the following are true, then articulatory representations may be quite useful for engineered speech recognizers:

Invariance of motor representations

Articulator configurations may be more invariant across contexts than acoustics. If so, such a representation would make it easier to classify sounds as they occur in continuous speech and varying environments.

For a given distinctive phonetic feature, there are a number of associated acoustic cues, each sufficient but not necessary for perception of the distinction (Liberman and Mattingly 1985). Articulation, on the other hand, is the basis for linguists' primary description of the feature, and provides a clear criterion for the distinction: e.g., tongue is low (section 1.9.3; section 1.3.5). Years ago, acoustics were easier to quantify than articulation (Harshman et al. 1977). More recently, tongue shapes were found to be less variable than previously supposed (Stone and Lundberg 1996).

The above argument regarding invariance is particularly compelling for various simple acoustic analyses. For example, the power spectral density at a particular frequency depends on both vocal-cord vibration and the vocal-tract resonances (section 1.12.3).

Reconciliation of audio and video

Converting auditory and visual inputs into a common articulatory representation might improve recognition by making it easier to compare them.

An analogy can be made to the combination of auditory and visual cues for localization. The optic tectum (part of the nervous system) of owls includes an array of neurons which encode position based on both types of input (Knudsen 1982) (Liberman and Mattingly 1985).

The McGurk effect is the classic demonstration of lipreading by listeners with normal hearing (section 1.3.1); its discoverers named the motor theory as a possible explanation (MacDonald and McGurk 1978).

1.1.1 Acoustic Cues

Some speech phenomena are readily observed in the acoustic domain. For example, voicing, the vibration of the vocal cords that distinguishes /z/ from /s/, greatly increases overall sound level, and introduces a periodic or harmonic component to speech. Stop consonants such as /d/ and /p/, which involve touching two articulators together, cause a rapid change in acoustics, as airflow is interrupted or allowed to resume. Sibilants have a concentrated band of energy at 4 kHz and above. This helps

explain why the sounds /s/ and /f/ are so hard to distinguish on the telephone, since telephones attenuate frequencies above 3600 Hz.

1.1.2 Describing Coarticulation in the Articulatory Domain

The speed of controlled articulator motion is limited (figure 1.7), causing coarticulation: the blending of successive sounds of speech. The blending effect is not limited to adjacent phones (acoustic units of speech), but can occur over an interval of up to five phones (Kent and Minifie 1977). Recordings are intelligible even at 400 words per second (Orr et al. 1965), but when speech is synthesized by splicing 20 phones per second together without blending, it is perceived as a buzz (Harris 1953) (Liberman et al. 1967). This indicates that coarticulation is necessary for intelligibility.

Coarticulation can be represented either in an acoustic analysis or by modeling articulator motion (the kinematic or motor representation). Which approach will ultimately work better is an open question. An acoustic account of coarticulation assumes that the smooth articulator motions translate into blending of acoustic parameters. Notably, some articulators change state much more quickly than others. For example, while the tongue moves gradually from position to position, the vocal cords may abruptly begin or stop vibrating. This voicing effect causes a dramatic change in acoustic signal strength. In speech synthesis, both types of blending have been used (section 1.12.3); for facial animation, kinematics is necessary (Parke and Waters 1996).

Under some circumstances, listeners can see what sound a speaker is about to make (Cathiard et al. 1992). This is due to anticipatory articulation: a speaker moves his or her mouth in preparation to make a sound.

In speech recognition, coarticulation has been modeled as a process of interpolation between acoustic targets (section 3.3.2) (Deng et al. 1992) or by enumerating and collecting separate statistics for sounds in different contexts (section 1.11.3).

In recognition with a motor representation (section 1.4), coarticulation can be modeled as movements are inferred from sound (section 3.3.6 and section 3.3.7) or in the back end recognizer architecture (section 1.10.1). The present thesis skips the inference step and uses kinematic measurements, directly observing motor variables subject to coarticulation (figure 1.7).

1.1.3 Motor-Space Reconciliation of Lipreading and Hearing Speech

Another motivation for representing speech in a motor space is to use that space to reconcile audio and video inputs for lipreading by computer. This is the so-called motor recoding approach, defined in section 1.7.1.

1.1.4 Sources of Kinematic Representations used in this Thesis

This thesis is a step toward answering the question of what information can be extracted using an inverse-problem, articulatory approach to recognition. Such approaches can be summarized by the the hypothesis that "We lipread by ear" (Paget 1930); this was originally proposed as a mechanism for human speech perception (section 1.2). A recognizer trained with articulatory (chap. 3) data might better describe the process of coarticulation, in which a speech sound is modified its phonetic context, than a recognizer that uses only speech sounds. Articulatory recognizers might also improve lipreading performance (chap. 2) by using articulation as a target representation for both sound and video channels. Related work has made significant progress toward recovering articulator motions from both sound and video. Although this thesis describes a new technique for recovery from video, it uses measured articulator positions to see how recognition might be improved with exact recovery from sound.

An assumption behind using articulatory data is that, eventually, a speech recognition system may be built that estimates articulator positions from sound and/or video. The motor theory of speech perception proposes that humans may use such representations when listening and watching speech. The present thesis does not present results for the process of human perception, but is concerned with the engineering of automatic recognizers.

The kinematic measurements used for recognition in this thesis were taken as indicated in figure 1.1. In chapter 2, a side-view camera and light were used to determine motion of regions near the upper and lower lip. Recognition experiments in chapter 3 used data for tongue, lip, and jaw motion collected by other researchers at the University of Wisconsin with an x-ray microbeam technique (section 3.4.1 and appendix E) (Westbury 1991). Other ways to measure tongue motion directly are listed in section 3.4.2.

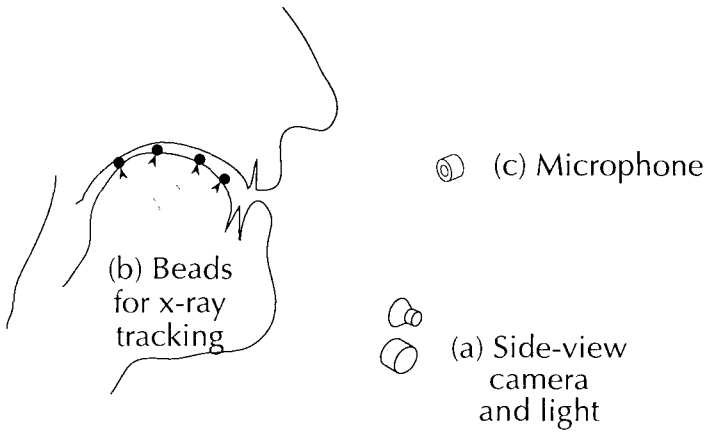


Figure 1.1: Sources of data used for the experiments of this thesis. Chapter 2: side-view lipreading (a) and sound (c). Chapter 3 and chapter 4: x-ray microbeam tracking (b) and sound (c). □

1.2 Articulatory Theories of Human Speech Perception

Many variations have been proposed of the basic idea that motor representations are used as an intermediate stage in speech perception. The schools of thought about perception that are most relevant to the present thesis are the gesture theory (section 1.2.1) and the motor theory (section 1.2.2).

1.2.1 Gesture Theory of Speech

The gesture theory proposes that speech originated in gestures (Rae 1862), and that articulator motions are recovered from sound in the listener's brain (Paget 1930). Since the recognizer of chapter 3 has access to direct articulatory measurements, no recovery from sound is used here. However, for the recognizer to be of practical benefit, direct measurement must be replaced by such a recovery step. The first claim of the gesture theory, regarding the evolutionary origins of speech in humans, has little to do with engineering automatic recognizers:

> Originally man expressed his ideas by gesture, but as he gesticulated with his hands, his tongue, lips and jaw unconsciously followed suit in a ridiculous fashion, "understudying" the action of the hands. (Paget 1930)

The motor theory (section 1.2.2) narrows the scope of the gesture theory by omitting the above claim of language origins. The common claim of the two theories is that

> the significant elements in human speech are the postures and gestures, rather than

> the sounds. The sounds only serve to indicate the postures and gestures which produced them. We lipread by ear. (Paget 1930)

1.2.2 Motor Theory of Human Speech Perception

In the study of speech perception, the motor theory proposes that the gesture units of articulatory phonology (section 1.2.4) are explicitly represented in the listener's brain. The motor theory has been defined as follows:

> The first claim of the motor theory... is that the objects of speech perception are the intended phonetic gestures of the speaker, represented in the brain as invariant motor commands... for example, 'tongue backing,' 'lip rounding,' and 'jaw raising'...
>
> The second claim of the theory is... [the link between speech perception and speech production] is not a learned association.... Rather, the link is innately specified.... (Liberman and Mattingly 1985)

Research results, described below, regarding language development in infants are part of the motivation for the second claim.

1.2.3 Motor Phonetics

Motor phonetics (Stetson 1988) differs from the approach of this thesis in that it uses the syllable as the fundamental organizational unit of speech analysis. There are so many possible syllables that recognizers are very rarely built which enumerate them all. Syllable-level analysis is discussed further in section 1.9.5. Motor phonetics describes how constraints on motion affect speech sounds; for example, limitations on acceleration and deceleration cause qualitative differences between fast and slow speech. This school of thought

1.2.4 Articulatory Phonology

Articulatory phonology (Browman and Goldstein 1992) describes the structure of speech systematically in terms of overlapping gestures. Unlike motor phonetics (section 1.2.3), it does not emphasize the syllable as a fundamental unit of speech. In contrast to the gesture theory (section 1.2.1) and motor theory (section 1.2.2), it does not require that the overlapping gestures be explicitly represented in a listeners' brain; it simply treats gestures as an expedient for phonology.

1.3 Evidence For and Against the Motor Theory

Various proposed versions of the motor theory have been disproven over the years, and correspondingly the theory has been revised.

- Phonetic perception precedes speech production in infant development, so if this level of perception requires motions to be inferred, that inference is not learned through speaking.

1.3.1 Lipreading with Normal Hearing

Lipreading by people with normal hearing and sight has been quantified and better understood over the past few decades; listeners use both auditory and visual information. Other researchers' results and a discussion of the number of degrees of freedom involved in human lipreading appear in section 2.2.1.

Lipreading in Noisy Environments

Since people without hearing loss understand speech quite well in quiet environments, the importance of lipreading is easier to demonstrate in a noisy environment (Sumby and Pollack 1954). This is certainly part of the cocktail-party effect, in which a person can selectively listen to one voice in a cacophony of other voices.

McGurk Effect

Severe contradictions between hearing and sight, such as mismatch of lip motion and soundtrack in dubbed foreign-language films, are obvious to the listener. However, subtle conflicts between a carefully chosen soundtrack and video result in subjects perceiving a sound other than that appearing on the soundtrack. For example, if someone hears /bɑː/ ("bah") while seeing a face saying /gɑː/ ("gah"), the listener will perceive the sound to be /dɑː/ ("dah"). This is known as the McGurk effect (McGurk and MacDonald 1976).

Speech sound can also influence visual perception of the face, in a reverse McGurk effect (Easton and Basala 1982). Like the McGurk effect, it is true both for artificial cases (such as /bɑː/ versus /gɑː/) and for actual words (Dekle et al. 1992).

The discoverers of the McGurk effect suggested the motor theory as its possible explanation (MacDonald and McGurk 1978).

Magnetoencephalography (MEG) and McGurk Effect

A recent magnetoencephalograph (Cohen 1972) study of the McGurk effect suggests that the perisylvian cortex plays a role in integrating audio and visual information (Sams and Levänen 1996).

1.3.2 Likely Mechanisms for Human Lipreading Ability

There is not yet agreement on how lipreading works in humans, and the present thesis does not claim to resolve the controversy. In section 1.7.1, a taxonomy of multimodal integration strategies for lipreading is listed (Robert-Ribes et al. 1996):

- Direct identification
- Separate identification
- Motor-space recoding
- Dominant recoding

One school of thought is that visual cues dominate perception of manner (e.g., /k/ versus /t/) of articulation, while auditory ones dominate perception of manner (e.g., /t/ versus /s/) (McGurk and MacDonald 1976) (Summerfield 1987).

A hybrid direct- and separate-identification model has been proposed (Massaro 1996), which has the disadvantage of being more complex than other models also consistent with the data (Robert-Ribes et al. 1996).

Noting that the auditory cortex is active during pure lipreading (i.e., without sound) (Calvert et al. 1997), some adhere to the dominant recoding theory, with the dominant modality being sound. This is not inconsistent with the above mentioned concept that some cues are more reliably conveyed by each modality. It simply claims that reconciliation of the two modalities occurs with a fundamentally auditory representation, which could be augmented with an indication of the reliability or certainty of each feature extracted from each modality.

1.3.3 Infant Language Development and Motor Representations

If motor representations are an essential part of human speech perception, it would be expected that such representations would be present in infant development at the time speech perception emerges. Babbling, the predecessor to speech, does not typically appear until seven to eight months

(van der Stelt and Koopmans–van Beinum 1986) (Holmgren et al. 1986) (Locke 1992). It is tempting to think that babbling teaches the infant the relation between motor acts and sound, and that this relation is the basis for learning low-level language processing. Yet research on language development shows that infants have started learning the phonetic contrasts specific to their native language by six months of age (Kuhl et al. 1992). Other studies (Eimas et al. 1987) have demonstrated infants' categorical perception (Harnad 1987) of coarticulated speech sounds. Either the articulation-acoustic relation is not needed for phonemic discrimination, or it is not associatively learned through babbling (Liberman and Whalen 2000) (Petitto and Marentette 1991) (Jordan and Rosenbaum 1989) (Sereno et al. 1987).

1.3.4 Magnetoencephalography (MEG) Studies

Magnetoencephalography (Cohen 1972) is a state-of-the-art technique for observing the timing and location of the brain's response to stimuli. It has been used for some time to observe the response of the cortex to speech (Hari 1991). For example, presenting subjects with complex sounds similar to speech elicits a right-hemisphere response if the sounds contain slow acoustic transitions, or a response in both hemispheres if the transitions are fast. In contrast, presentation of syllables alone specifically activates the left hemisphere. The short latency of these responses suggests that they precede the brain's mechanisms for attention (Shtyrov et al. 2000). That the left hemisphere processes many aspects of language was already known; the above results indicate that speech-specific low-level regions of the left hemisphere reject acoustically similar nonspeech sounds. Another study (Imaizumi et al. 1998) found that the right hemisphere predominated in word discrimination by intonation, while the left hemisphere dominated in discrimination by a phonemic minimal pair.

Taken together, these results are consistent with the left hemisphere processing vocal tract, rather than vocal cord, (section 1.12.3) characteristics of speech. They do not indicate whether this processing is based on an acoustic (formant) representation, or a motor (geometric) representation.

A recent review of evidence, including brain imaging and aphasias, argues against the use of articulatory representations in human speech perception (Coleman 1998). The argument is based largely on the proximity of different types of processing in the cortex, and on the different contexts in which particular cortical regions are active. It also cites the functions during which certain structures are active: for example, Broca's area (Broca 1865) (Kandel et al. 1991) has a role in recognition of sentence structure, but not in recognition of individual words or spontaneous interjections.

1.3.5 Language Instruction

Foreign-language instructors routinely use articulatory, rather than acoustic, explanations to describe sounds not present in a student's native tongue. For example, an advanced English-speaking student of Spanish might be instructed to make the intervocalic /β/ sound (as in Spanish "ave") like an English /v/, but without touching the bottom lip to the teeth.

1.4 Articulatory Recognition Proposed

A paper on the prospect of artificial intelligence (MacKay 1951) described two different approaches to designing thinking machines: reception and replication. The reception approach would involve comparing incoming stimuli to templates of idealized objects; replication was described as follows:

> Let us... consider the way in which a blindfold man might seek to recognize a solid triangular figure, by moving his finger around the outline.... To the blindfold man, the concept of triangularity is invariably related with and can be defined by the sequence of elementary responses necessary in the act of replicating the outline of the triangle. Let us generalize this approach to the problem of recognition, and consider now an artefact whose response to incoming stimuli of any kind is an *act of replication*, in some formal sense, of the stimuli received. (MacKay 1951)

The inverse-problem, articulatory approach to automatic speech recognition, an example of the replication approach, had already been proposed (Dudley 1940). Over the years, a number of researchers have contributed to this eventual goal, as described in section 3.3.

1.5 State-of-the-Art Automatic Speech Recognition

Many speech recognition systems are commercially available. They do not typically incorporate articulatory representations, and they have a number of shortcomings. For best performance, large-vocabulary systems require the user to wear a microphone headset. Accuracy degrades considerably with desktop microphones in an office environment.

Another problem with large-vocabulary dictation systems is the correction interface. Text appears on the screen after a delay, so the user may have spoken several more words by the time an error is

evident. Even if the system flags the word as not understood, it is likely to get it wrong again even if the user repeats it very precisely.

1.6 Measuring Error Rates

An isolated-word recognizer is assessed by comparing each word label it guesses to the corresponding word in the true text of what was spoken; for such a recognizer, the error rate is simply the number of wrong guesses divided by the number N of total words. The chance rate, achieved by choosing words randomly, is $(N-1)/N$. In continuous speech, the recognizer may miss a word (a *deletion error*) or may output an extra word (an *insertion error*). It may output the right word in response to the wrong interval of speech; however, if the texts match such an error is not normally counted against the recognizer. The third type of mistake included in the error rate is a *substitution error.*

Output from a continuous recognizer is evaluated by finding an optimal *alignment* between the guessed text γ and the true text τ that minimizes the total number of errors. An alignment is a sequence of pairs of matching words, one per pair coming from each text; it is generated using standard algorithms for approximate string matching (Sankoff and Kruskal 1983). The error rate E is then the total number of errors (deletion D, substitution S, and insertion I) divided by the length of the true text:

$$E = \frac{D(\tau,\gamma) + S(\tau,\gamma) + I(\tau,\gamma)}{\|\tau\|}$$

In general, a continuous recognizer's error rate may exceed 100%. For example, the recognizer might output too many words, all of them wrong.

Because of the alignment process, a recognizer that outputs the right total number of words but chooses them randomly will have a lower error rate than $(N-1)/N$. Predicting chance performance is nontrivial and has been analyzed for the special case of matching string lengths and counting each substitution error two errors (insertion plus deletion) (Deken 1979).

In chapter 3, *error rates* are used instead of *percent correct* to evaluate recognition performance. This choice is motivated by the fact that the recognition task is continuous speech: i.e., the recognizer is presented with speech not previously segmented into linguistic units. In contrast, in chapter 2, each input example consists of an isolated audio or video sample of a single word.

1.6.1 Error Rates as the Only Metric: Shortcomings and Advantages

In assessing recognition performance for conversations, it is worth noting that many of the most common words and phrases are vernacular interjections, which might reasonably be omitted in a transcript. By concentrating on these common words, the overall error rate may be reduced (Byrne et al. 1998), but the intelligibility of the resulting transcript is not correspondingly improved.

Error rates may be misleading. In the above-described case of conversations, it can be useful to think in terms of the information lost due to errors. The true prior distribution of utterances is needed to calculate information provided by the recognizer and entropy of the recognition task. Unfortunately, this distribution is not generally available and must be approximated by the recognizer. A theoretical framework exists for measuring the information provided independently by grammar and acoustic models, as well as by their joint operation in a recognizer (Ferretti et al. 1990). Recognizer parameters may be chosen during training to maximize mutual information (Bahl et al. 1986) (Rabiner and Juang 1993). Still, the standard for recognizer assessment is error rate.

1.6.2 Task Specificity

In general, speech recognizers have always had problems generalizing from the specific task for which they were developed. It is now recognized that definitive assessment of recognizers requires that data be split into training, development testing, and evaluation testing sets. In the early years of ASR, data would sometimes not even be split into training and testing sets (Jelinek 1996). Since this type of experimental design revealed nothing about generalization, critics suggested that these early recognizers generalized poorly (Pierce 1969).

A true evaluation set is used only once, to simultaneously compare the performance of different recognizers. The development testing set may be used repeatedly while making modifications to the recognizer architecture, while the training set is used to fit model parameters. The distinction between architecture and parameters is not absolute. Global parameters such as the grammar weight (section 3.11.2), fewer in number than model parameters and harder to optimize, may require that the development test be further subdivided for cross-validation.

All of the above considerations in splitting the data set miss a larger point, which is that existing corpora are often homogeneous—consisting, for example, only of informal telephone conversations (Godfrey et al. 1992), or of people reading from the *Wall Street Journal*, or of newscasts. A recognizer trained on the latter two cases would have a great deal of trouble on conversations; a telephone

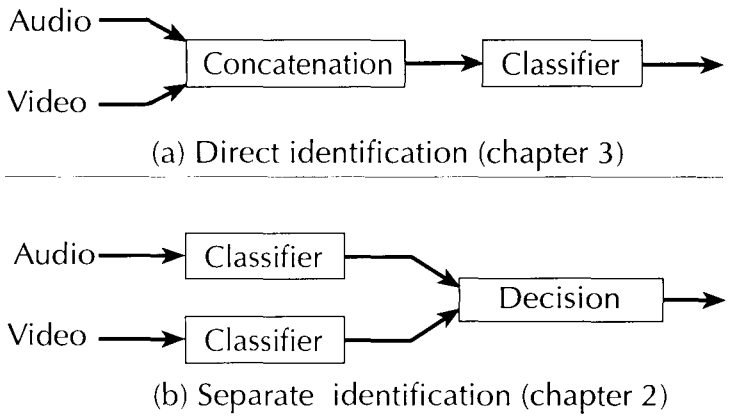


Figure 1.2: Different approaches to multimodal integration (Robert-Ribes et al. 1996). Chapter 3 uses the direct identification approach (a) to merge audio and direct articulatory measurements, and the analysis of chapter 2 is closely related to separate identification (b). □

conversation transcriber would have trouble transcribing a discussion recorded from a single microphone in the area, rather than close-talking microphones on the participants.

1.7 Relation between Lipreading and Articulatory Recognition

1.7.1 Approaches to Multimodal Input Integration

There are several ways to combine different types of input in a speech recognizer. For lipreading, the approaches have been categorized as follows (figure 1.2; figure 1.3) (Robert-Ribes et al. 1996), but the framework also applies to other multimodal problems such as joint acoustic-articulatory recognition.

- *Direct identification* (figure 1.2(a)), in which audio and video inputs are presented to the recognizer without transformation to a common space. This is the approach used in chapter 3.

- *Separate identification* (figure 1.2(b)), in which audio and video signals are separately classified as belonging to a particular linguistic unit (e.g., phoneme), and classifier outputs are combined. In general, the classifier's outputs may take on continuous values. This is the approach used, for a very small data set, in chapter 2.

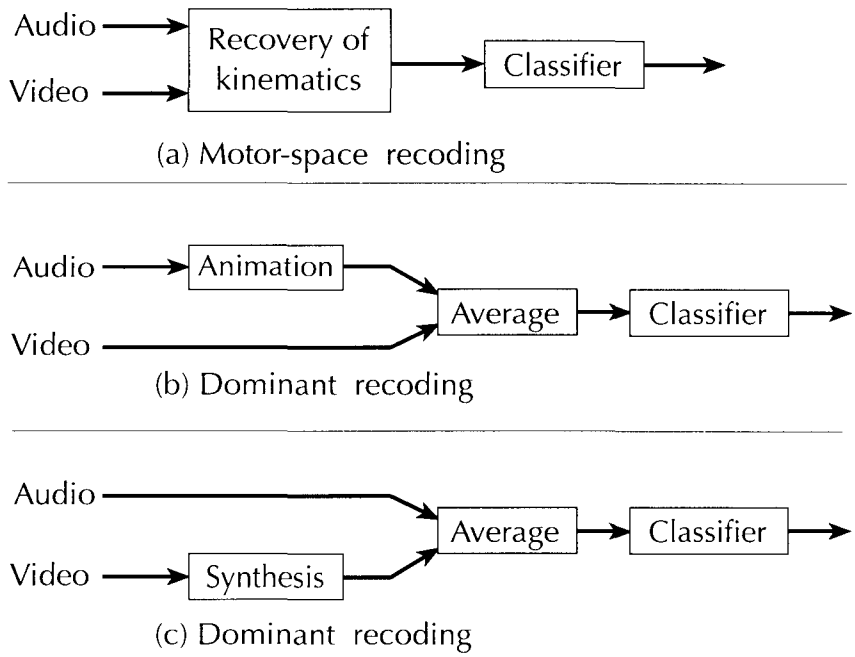


Figure 1.3: Additional approaches to multimodal integration (Robert-Ribes et al. 1996), not implemented in this thesis, but illustrated here for discussion purposes. □

- *Motor-space recoding* (figure 1.3(a)), in which both audio and video are used to derive articulatory parameters such as distance between the lips. As described below, this approach is the long-term goal of this thesis.

- *Dominant recoding* (figure 1.3(b); figure 1.3(c)), in which one modality is considered more important for recognition, and the second modality is converted to a similar representation. An example would be predicting cepstral coefficients (section 1.10.2) based on video.

Direct and separate identification are the most widely used lipreading strategies because it is easier to build recognizers based on such principles than on motor-space or dominant recoding. To date, no approach has been conclusively demonstrated to outperform the others, so researchers' choices have been motivated by theoretical considerations such as simplicity or elegance. An advantage of either recoding technique over separate identification is the measurability of the intermediate representation: in other words, motor recoding derives quantities, such as speed of jaw opening, that are easier to objectively define and measure than abstract category scores. The motor representation, as opposed to a category score, retains more information about the original signal, since articulatory parameters can reconstruct the original sound (speech synthesis) or image (facial animation) more

accurately. The retained information includes the detailed timing of individual speech sounds; in separate identification, the timing is discarded before the audio and video are merged. Direct identification retains timing information, but makes it difficult to model the correlations between the two (audio- and video-derived) parameter sets.

The above approaches can be thought of as part of a taxonomy, in which three successive questions rule out alternatives (Robert-Ribes et al. 1996):

- Are audio and visual stimuli converted to a single representation in which distances can be computed? If not, the strategy is direct identification
- Is the common representation language-specific, or based on abstract linguistic processing? If so, separate identification is being used
- The final decision is between motor space recoding and dominant recoding, as defined above

Approach of this Thesis

The work of this thesis represents a few steps toward the ultimate goal of motor-space recoding. Techniques are described that extract motion information from video signals (chap. 2), and results are reported for augmenting recognition from sound with motion information (chap. 3). Merging the two components (motion from video; recognition from motion plus sound) is outside the scope of this thesis. The preliminary lipreading results of chapter 2 were based on separate identification, and the merging of sound and articulation in chapter 3 is a hybrid of motor recoding and direct identification.

A previous project implemented the dominant-recoding approach using a neural network, which was presented with raw image data (an array of pixel values) (Yuhas et al. 1989). The network predicted the sound spectrum from the image, and this prediction was averaged with the actual spectrum.

Coarticulation, the process by which spoken sounds are dramatically changed by preceding and following sounds, seems to have more to do with getting from one mouth configuration to another (articulatory smoothing) than the similarity of the resulting sounds (acoustic smoothing) (Stetson 1988).

1.8 Timescales of a Recorded Word

Many ideas about speech are inspired by the slow motion of the articulators compared to the oscillations of the acoustic waveform. In figure 1.4, different timescales of speech have been extracted with a Haar wavelet decomposition. The comblike shape in the original waveform and finest scales of detail has a period of about 9 ms and represents the vocal cord vibration. For voiced sounds, the fundamental frequency is the reciprocal of this period. The plosives /d/ and /t/ are also quite distinct at the finest levels of detail.

In figure 1.4, each time series is normalized. Without normalization, the original signal equals the following sum of coefficients:

$$s(t) = a_{10}(t) + \sum_{i=1}^{10} d_i(t)$$

Two vocal cord (fundamental frequency) cycles from the first syllable of "dormitory" appear in figure 1.5. The oscillations within each individual cycle are due to the resonances of the vocal tract (section 1.12.3).

The power spectrum for the voiced interval of figure 1.5 appears in figure 1.6, indicating which frequencies are most prominent. Harmonics of vocal cord vibration appear as a comb in the spectrum. The local peaks around 500 and 1000 Hz are near formant frequencies—resonances of the vocal tract.

An automatic procedure estimated the fundamental frequency of figure 1.6 to be 111 Hz. The spectrum was generated with a 93 ms (i.e., wider) interval centered at the same time as figure 1.5, Hanning windowed, and 2^{11}-point fast Fourier transformed.

In figure 1.7, sound and articulator movements are plotted for the same word; the data are from the set used for chapter 3.

One way of thinking about the relation between the two scales is that the vocal cords provide a carrier signal, which is then modulated by the changing resonances due to articulator motion (Dudley 1940).

1.9 Linguistic Units of Speech

Linguistics uses different units for the various timescales and levels of abstraction of speech; these include phonemes, allophones, gestures, distinctive features, syllables, words, and sentences. As

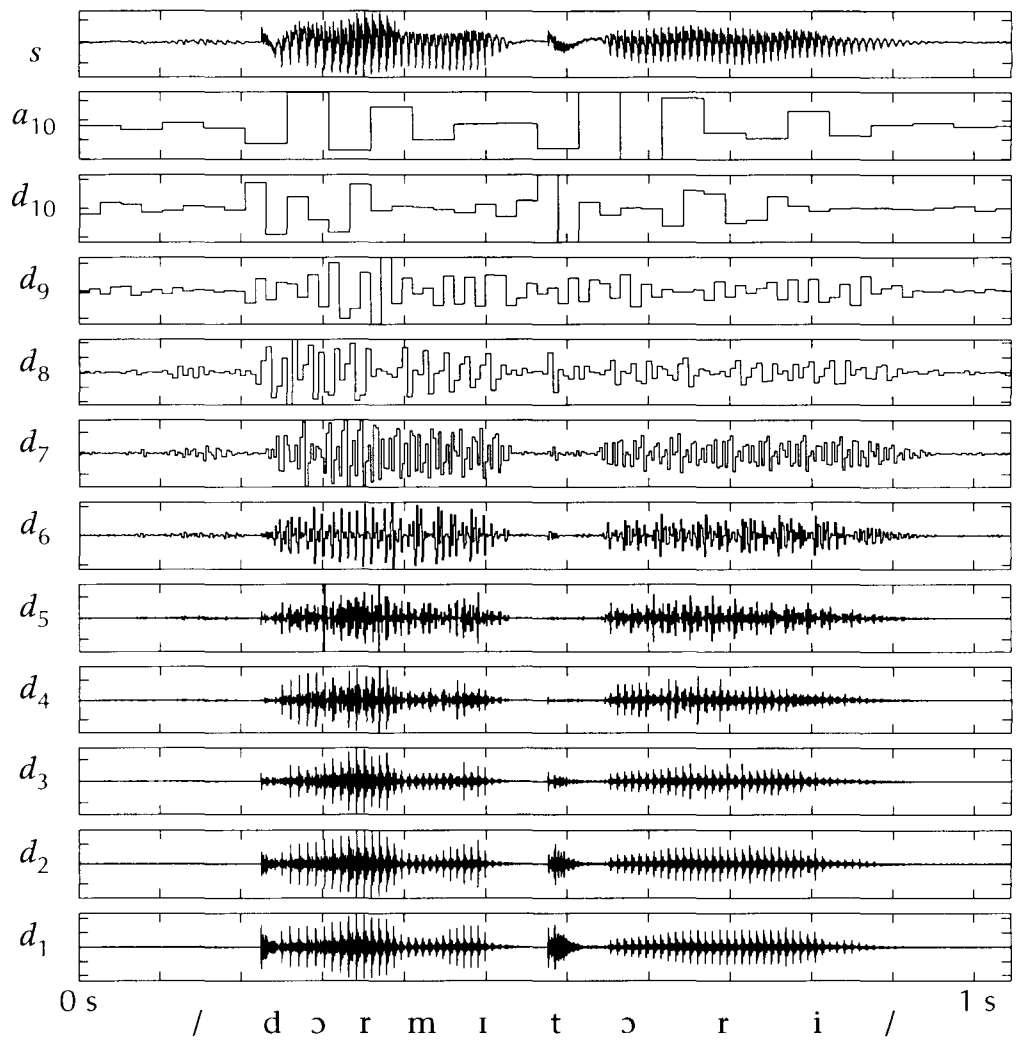


Figure 1.4: Haar wavelet decomposition showing the different timescales of a recorded word; word spoken was "dormitory." Even the shortest (fastest) timescales contain phonetically relevant information: specifically, plosive bursts /d/ and /t/. The articulator motions of figure 1.7, in contrast, include only much longer timescales. □

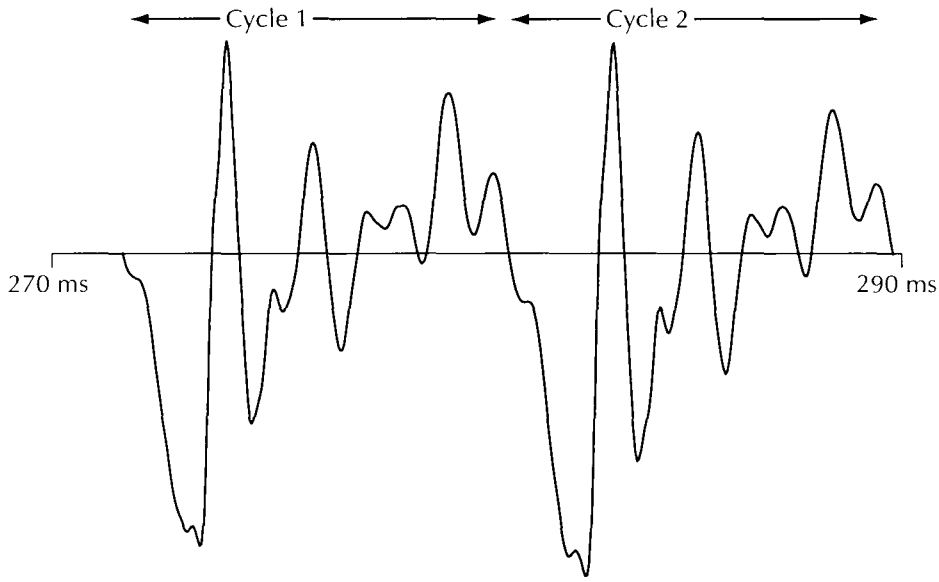


Figure 1.5: Sound waveform for two cycles of vocal-cord vibration; spectrum appears in figure 1.6. The higher-frequency oscillations within each of the two cycles are due to the resonances of the vocal tract. Recognition generally requires an acoustic analysis of these signals; the pure waveform representation depicted here is extremely hard to classify. □

described below, although phonemes and allophones use the same symbols, they represent different levels of abstraction and are distinguished in writing by the delimiters between which they appear. The variation between allophones and the number of possible syllables present challenges in recognition.

1.9.1 Phonemes

Phonemes are units in a particular language that differentiate words from each other. They are defined by analyzing semantically-distinct pairs of words that differ by a single minimal pair of sounds. Each language has a particular set of phonemes, with distinct phonemes of one language being merged in another. By convention, phonemes are enclosed in slash symbols—e.g., /k/.

1.9.2 Allophones

An allophone is an acoustically but not semantically distinct version of a phoneme, presenting additional variability to be addressed in a speech recognizer. A word pronounced with the right phoneme

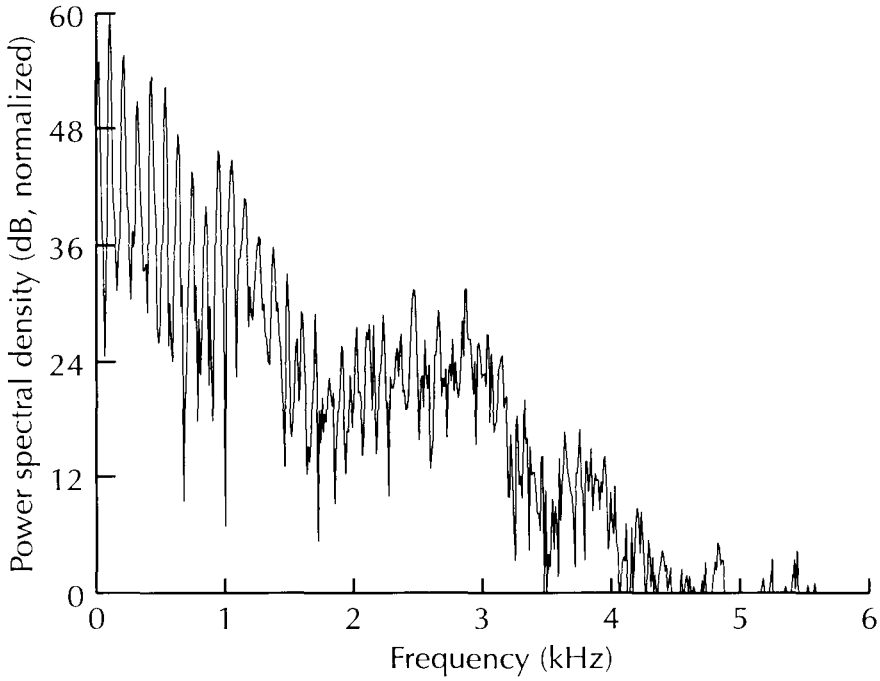


Figure 1.6: Speech spectrum: an easier representation for recognition than figure 1.5. In a typical recognizer, the cepstrum (section 1.10.2) is used instead of the spectrum. This plot is calculated over a slightly longer time window centered at the same time as figure 1.5. The regularly-spaced peaks are vocal-cord harmonics, and elevated parts of the overall spectral shape are formants. □

but the wrong allophone is generally intelligible but unnatural-sounding. The decision of which allophone to use is generally based on the surrounding sounds. The English phonemes /r/ and /l/ are allophones of a single phoneme in Japanese. Whether a consonant will be aspirated—followed by a brief forceful exhalation—is determined by context in English, making it an allophonic distinction. In Hindi, aspirated and unaspirated versions of the same stop consonant are different phonemes. Allophones are conventionally written with bracket symbols—e.g., [p^h].

1.9.3 Distinctive Features

In linguistic analysis, sounds are described by sets of distinctive features. Initially, these sets were largely acoustic (Jakobson et al. 1952) (Trubetozky 1939), but more recent systems of distinctive features are based on articulation (Liberman et al. 1967) (Chomsky and Halle 1968).

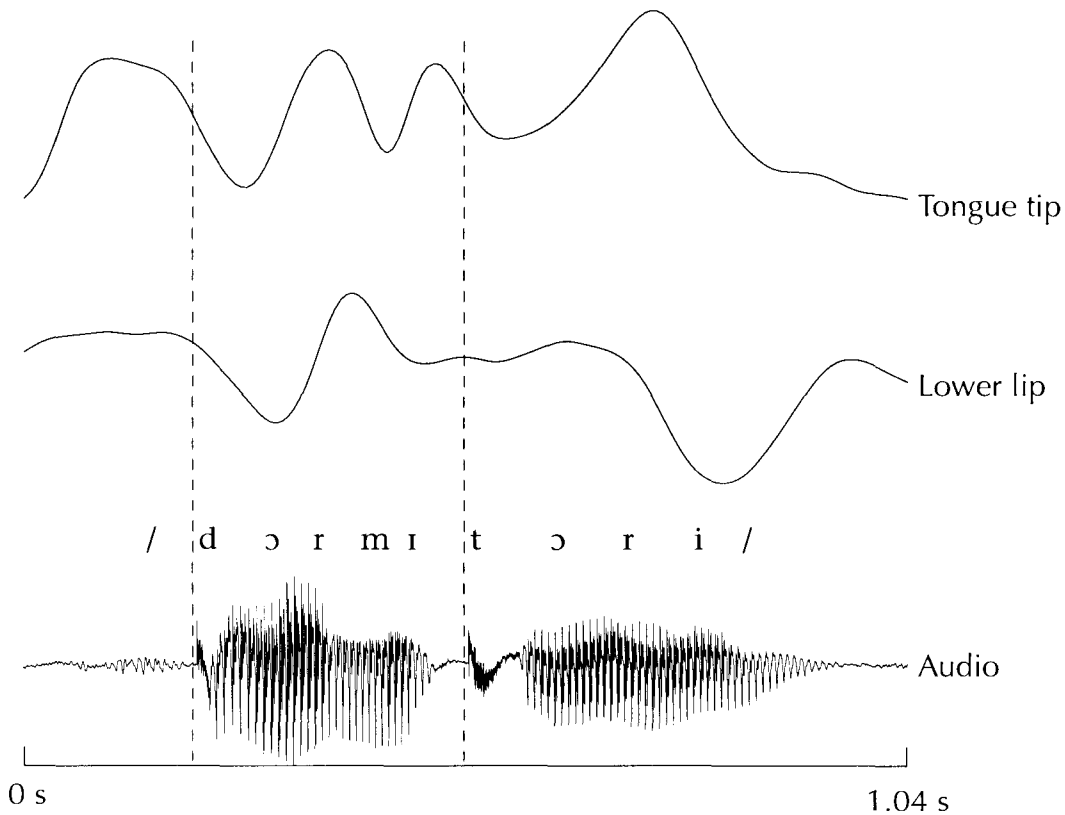


Figure 1.7: Timescales of articulator movement and sound. The lip motion contains a great deal of phonetic information—especially when combined with tongue motion and detection of voicing. Motion occurs on a similar timescale to the sequence of phones, much slower than the fastest oscillations in sound. Speaker is saying "dormitory." □

1.9.4 Gestures

Gestures are very similar to articulatory features—a gesture is essentially the realization of an articulatory feature (such as used in chapter 3) for some interval. The term is generally used by those concerned with the dynamic process of coarticulation, the relative timing and interaction of gesture sets for successive phonemes (section 1.2.3; section 1.2.4); or by those asserting a *fundamental link* between perception of hand gestures and perception of speech (section 1.2.1).

1.9.5 Syllables

Some analyses of speech—notably motor phonetics (section 1.2.3)—emphasize the syllable as an organizational unit. In a syllable, there are many possible combinations of releasing (initial) consonants, vocalic (central) sounds, and arresting (final) consonants. In English, there are at least 61 releasing consonant sequences (e.g., /skr/ as in "scratch") and at least 114 arresting consonant sequences (e.g., /ts/ as in "rats") (Stetson 1988). As a result, there are an unwieldy number of distinct syllables of English, so the present thesis avoids explicit syllable modeling. For the same reason, few syllable-level recognizers have been developed by other researchers.

1.9.6 Words

For speech recognition purposes, it is important to recognize that unlike text, in which words are separated, spoken words are not acoustically disjoint. They flow together due to coarticulation. Also, the pronunciations listed in a dictionary do not capture the ways words are actually realized in speech.

1.9.7 Importance of Pitch

In the following discussion, "pitch" is used as shorthand for "fundamental frequency of vocal fold vibration." The claims do not necessarily apply to musical or perceptual notions of pitch.

For tonal languages such as Chinese, words with the same phonemic pronunciation but different pitch have distinct semantics. In such cases pitch is retained for recognition (Yang et al. 1988).

In other languages, the role of pitch is less significant in distinguishing words. Nevertheless, it may indicate a stressed syllable, and phonemically identical words may differ only in where the stress falls. In English pitch and loudness are both used to indicate stress, while Japanese only uses pitch. Information about syllable stress can improve recognition (Zue 1985), but most present-day recognizers discard pitch information by using a truncated cepstrum (section 1.10.2). In word recognition by humans, prosody activates brain areas distinct from those activated by phonemes (Imaizumi et al. 1998).

1.9.8 Sentences

Sentence structure is beyond the scope of this thesis, and conversation analysis (Zue and Glass 2000) even more so. The work of chapter 2 involved only isolated words. Where grammar models were

used, they used only frequency and transition statistics: chapter 3 used a unigram model, while appendix B and chapter 4 used a trigram model (section 1.11.4).

1.10 Front-End Algorithms in Conventional State-of-the-Art Recognizers

1.10.1 Separation into Front and Back Ends

Speech recognizers typically have two components operating in series: a front end that transforms sensor data (e.g., audio waveforms) into a set of continuous variables changing over time; and a back end that guesses what sequence of discrete textual units (e.g., words) was spoken. In the side-view lipreader of chapter 2, the video front end is a new processing pipeline (figure 2.7) and the back ends are conventional Hidden Markov Models (section 1.11.2) and maximum-likelihood Gaussian classifiers. The steps used in chapter 3 appear in figure 3.6. An optimization approach to selecting operating parameters of the front-end processor is the subject of appendix B.

1.10.2 Acoustic Analysis Using the Cepstrum

The cepstrum, used in both chapter 2 and chapter 3, is perhaps the most widely-used acoustic front end for speech recognition. It is generated, as described below, by inverse Fourier transforming the logarithm of a power spectrum (Bogert et al. 1963) (Rabiner and Juang 1993). One possible motivation for use of the cepstrum is that it is well suited to encoding the characteristic resonances (formants) of vowels; however, recognizers based on a cepstrum front end may perform just as well on consonants as on vowels (chap. 4).

Basic Cepstral Transformation

The cepstrum is based on the concept of homomorphic deconvolution—a technique that can, under certain circumstances, separate a signal into source and filter components. Consider a signal y that is generated by linear, time-invariant filtering of a source signal x:

$$y(t) = \int_{-\infty}^{\infty} h(\tau)x(t-\tau)d\tau$$

Equation 1.1: Source-filter model of speech, using a linear, time-invariant approximation □

In the classic model of vowel synthesis (section 1.12.3) h would represent the impulse response of the vocal tract. It would be determined in part by the shape of the mouth cavity, which in turn would be determined by tongue position. In practice, h would change over time, so this time-invariant analysis is only an approximation (see section 1.10.2). For vowel synthesis, x would be a series of pulses due to the vocal cords.

The filtering operation can be represented in the frequency domain as follows:

$$Y(\omega) = H(\omega)X(\omega)$$

where Y, H, and X are the Fourier transforms of y, h, and x. For example,

$$Y(\omega) = \int_{-\infty}^{\infty} x(\tau)e^{i\omega\tau}d\tau$$

In speech recognition, the signal y is typically subdivided into short time intervals (for example, 25 ms; see section 1.10.2), and phase information is discarded within those intervals. In the frequency domain, the phase of Y at frequency ω is given by the angle of the complex number $Y(\omega)$; the power spectrum $|Y(\omega)|^2$ discards phase by taking the magnitude:

$$|Y(\omega)|^2 = |H(\omega)|^2 |X(\omega)|^2$$

In homomorphic signal processing, filtering is represented by adding two quantities at each frequency, rather than multiplying. This change of representation is accomplished by taking the logarithm of each side of the above equation:

$$\log |Y(\omega)|^2 = \log \left(|H(\omega)|^2 |X(\omega)|^2 \right)$$

$$= \log |H(\omega)|^2 + \log |X(\omega)|^2$$

The final step in taking the cepstrum is to take an inverse Fourier transform of the log power spectrum $\log |Y(\omega)|^2$.

$$C(q) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \log \left| \int_{-\infty}^{\infty} y(\tau) e^{i\omega\tau} d\tau \right|^2 e^{-i\omega q} d\omega$$

Equation 1.2: Definition of the cepstrum C of a signal y □

Because the Fourier transform is a linear operation, the cepstrum of y will be the sum of the cepstra of h and x:

$$\int_{-\infty}^{\infty} \left(\log |H(\omega)|^2 + \log |X(\omega)|^2 \right) e^{-i\omega q} d\omega = \int_{-\infty}^{\infty} \log |H(\omega)|^2 e^{-i\omega q} d\omega + \int_{-\infty}^{\infty} \log |X(\omega)|^2 e^{-i\omega q} d\omega$$

Computation of Discrete Cepstrum

A typical signal-processing application will start with a sequence of sampled data and make use of the discrete cepstrum rather than the continuous form given above. The fast Fourier transform (FFT) is generally used as one step of the discrete cepstrum's implementation.

In speech recognition, as in many applications, the acoustic signal is nonstationary: its frequency spectrum changes over time. The typical way of working around this problem is windowing, which involves picking a time scale on which the spectrum seems relatively constant, and analyzing a series of time intervals of the appropriate size. The precise definition of a time-varying power spectrum is a research area unto itself (Cohen 1995) (Loughlin et al. 1994) (Fonollosa 1996); for this thesis, the conventional windowing and time-derivative (section 1.10.2) approaches are used, despite their limitations.

Truncation of Cepstrum

In speech recognition, there is a tradeoff between (1) having enough input parameters (typically called "features" (Duda and Hart 1973)) to distinguish different categories, and (2) keeping the number of model parameters low enough that the latter can be trained from the limited set of examples. The present thesis uses HMMs with continuous emission densities and diagonal covariance matrices (section 1.11.2), in which the number of model parameters is directly proportional to the number of input parameters.

To reduce input dimensionality, the recognizer uses only the first few coefficients of the cepstrum. These coefficients correspond to the smallest so-called quefrencies (see glossary) q in equation 1.2. For example, in chapter 3, the lowest 12 cepstral coefficients (not counting zero quefrency) are used:

$$C(n\tau) : 1 \leq n \leq 12; \tau = 46.0\mu\text{s}$$

The use of 12 coefficients is a standard practice, motivated by their ability to retain the formant frequencies of vowels, while removing from the spectrum the detailed harmonic information which signifies pitch (Rabiner and Juang 1993).

Spectral Preemphasis

Speech, like many signals of interest, has an approximately $1/f$ frequency spectrum. Recognition performance is improved when the spectrum is equalized by a preemphasis step, implemented as follows:

$$\dot{Y}(\omega) = Y(\omega)\omega^{\kappa}$$

Typically, in the above equation, $\kappa \approx 1$; for example, $\kappa = 0.97$ was used for chapter 3.

Mel-Frequency Warp

The Mel-frequency warp is an optional processing step used both here and in conventional recognizers. Inspired by the psychophysics of human pitch perception, it involves modifying the power spectrum to conform to the experimentally determined Mel scale of pitch (Stevens and Volkmann 1940) (Rabiner and Juang 1993), and it improves recognizer performance (Davis and Mermelstein 1980). It might simply be a coincidence that the Mel warp improves the performance of automatic recognizers, or there may be some undiscovered fundamental principle responsible for both the recognition improvement and human pitch perception. A recent project tested whether the Mel scale was the optimal choice from a large class of monotonic frequency warps; the optimal warps looked qualitatively like the Mel scale (Kamm et al. 1997).

Energy Terms

Concatenating energy terms to the front-end feature vector also improves recognition (Nocerino et al. 1985). These terms are the logarithm of the sum of squares of signal values within an interval (Young et al. 1997):

$$E(y, t, w) = \sum_{\tau=t-w}^{t+w} y^2(\tau)$$

The intervals are the same as those used for computation of the discrete cepstrum.

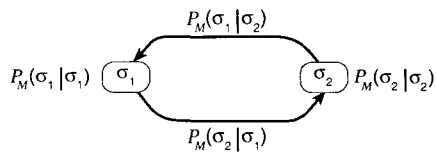


Figure 1.8: Example Markov model (MM) with two states. This model generates sequences of symbols such as "$\sigma_2\sigma_1\sigma_1\sigma_2\sigma_1$..." In a hidden Markov model, either state can produce either symbol, and P_M is split into P_T and P_E (figure 1.9). □

Time Derivatives

In front-end processing, the dynamic acoustic aspects of speech are represented by adding the time derivative of the spectrum or cepstrum. These features represent dynamics at the time scale of articulator motion—other than vocal cord vibration, several cycles of which can appear in each time interval of spectral analysis. The utility of time derivatives for automatic recognition was demonstrated decades ago (Fry and Denes 1958) (Denes and Matthews 1960), and as back-end architectures have evolved they have remained useful (Furui 1986). First and second time derivatives are computed, via numerical differentiation, for both cepstral coefficients and the energy term (Rabiner and Juang 1993).

Complete Feature Vector

The complete feature vector for a typical front end consists of 12 cepstral coefficients, one energy term, 13 first derivatives with respect to time, and 13 second derivatives, for a total of 39 dimensions. This parameterization was used throughout the present thesis.

1.11 Conventional Back-End Algorithms

1.11.1 Markov Models

Markov models are stochastic state machines that generate strings. They are used for higher-level processing in speech recognition—grammar modeling—in conventional recognizers such as the one analyzed in chapter 4. Hidden Markov models (section 1.11.2) are used for bottom-up processing. When used for recognition, instead of generation, strings are scored based on how likely it would have been that the model would have generated them. They have a set of states, one per symbol,

and a set of transition probabilities from state to state. Since they are stochastic, each generated string is chosen randomly.

In automated musical composition, Markov models have been used as generators, starting in the 1950s (Pinkerton 1956) (Brooks et al. 1957) and continuing over the decades (Ames 1989). Computer music has also used hierarchical models (Polansky et al. 1987). When text models are used as generators, the result is generally neither syntactically correct nor semantically coherent, but sounds vaguely plausible. An example, using models trained on the text of this dissertation, appears in appendix F.

1.11.2 Hidden Markov Models

Hidden Markov models (HMMs) were used to recognize audio in chapter 2 and both audio and articulatory data in chapter 3; HMMs are ubiquitous in speech recognition (Jelinek 1998b). One HMM represents each unit of speech (each word in chapter 2; each monophone in chapter 3). The recognition task helps determine what level of abstraction should be used for the units: for example, words versus individual phone-like units.

The so-called three basic problems (Rabiner and Juang 1993) for recognition and training with HMMs are the following:

1. The *forward problem* is useful for selecting among a small number of possible recognizer outputs (hypotheses). Given a model M and an observation sequence v (section 1.11.2), what is the likelihood that the model would produce the sequence? For example, consider the problem of recognizing isolated digits. The recognizer would use a set of models $M_i : 0 \leq i \leq 9$ for the digits. For an input utterance v, a score $p(v|M_i)$ would be computed for each possible digit i, and the recognizer would guess the digit i whose model had the highest score. Section 1.11.2 describes the computation in more detail.

2. *Viterbi decoding* is used with continuous speech, for which the number of possible output texts grows exponentially as the input sound gets longer. To recognize a sequence of digits, a composite model C would be formed including each of the digit models in a loopback configuration (section 1.11.2). Viterbi decoding gives the most likely state sequence λ through the composite model; the sequence of digits whose models λ passes through is the recognizer's output (section 1.11.2).

3. *Parameter estimation* is the problem of training a model from data. Given a model M and

a set of observation sequences $\{v\}$, how can M be modified to maximize the likelihood of it producing $\{v\}$? Assuming that the training set $\{v\}$ is large enough, the model's architecture and stochastic constraints prevent M from assigning arbitrarily high likelihood to all training examples, making it possible for training to converge.

Discrete Acoustic Input for HMMs

To simplify the following review of HMMs, sound recordings will be considered transformed into acoustic symbol sequences. These acoustic symbols will generally not correspond to a linguistic unit such as a phoneme, although they have occasionally been used as *fenones*, a substitute for allophones (Bakis 1974) (Jelinek 1998b). Usually, they represent purely acoustic categories, and describe a given interval of the sound recording (typically 25 ms long) with a single integer. The number of symbols would typically be of order 256. The symbols are determined for each interval by computing the cepstrum (section 1.10.2) and applying vector quantization (Jelinek 1998b); the latter is outside the scope of this thesis. Intervals are typically overlapping, each starting 10 ms after the start of the previous one.

The discrete acoustic representation described above reduces a sampled waveform, which might have a 16-bit sample value every 45 μs, to an 8-bit integer every 10 ms—a data compression ratio of 444:1. State-of-the-art recognizers have replaced discrete acoustic representations and discrete-emission HMMs with a continuous representation and HMMs having continuous emission densities (section 1.11.2).

Forward Problem for Isolated-Digit Models

Setting aside the question of how to train models from speech data (section 1.11.2), assume that an HMM M_i has been appropriately trained for each of the ten digits i. Each model includes a set of of states Q. Assume that the states are numbered with state 1 being the start state, and state $|Q|$ being the end state; the model progresses through the states during the course of the word.

For each pair of states q_i, q_j, a transition probability $P_T(q_j|q_i)$ is defined, giving the probability that the model will go to state q_j at time step $t+1$ if it was in state q_i at time step t. The probabilities P_T can be thought of as a square matrix of values, with the number of rows and columns each equal to the number of states. For example:

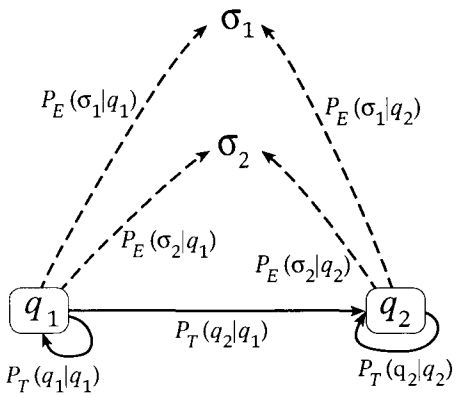


Figure 1.9: Example of a hidden Markov model (HMM) with two states q_1 and q_2 and two possible observations σ_1 and σ_2. In this example, the states have feed-forward connections, and any symbol sequence can be generated, though perhaps with a low probability, by any state sequence. □

$$\begin{bmatrix} P_T(q_1|q_1) & 0 \\ P_T(q_2|q_1) & 1 \end{bmatrix}$$

The entry for $P_T(q_1|q_2)$ is zero in this case because the model of figure 1.9 is feed-forward, as are typical speech-recognition acoustic models. If this entry were nonzero, the model would be referred to as ergodic (see glossary).

Time t is discretized, with each step moving forward one acoustic-analysis interval (10 ms; section 1.11.2). Because P_T represents conditional probabilities, the following stochastic constraints apply:

$$0 \leq P_T(q_2|q_1) \leq 1$$

$$\sum_{q_2} P_T(q_2|q_1) = 1$$

For each state $q \in Q$ and each possible acoustic symbol σ (section 1.11.2), an emission probability $P_E(\sigma|q)$ is defined. Like P_T, P_E can be thought of as a matrix with as many rows as there are input symbols, and as many columns as model states; e.g.:

$$\begin{bmatrix} P_E(\sigma_1|q_1) & P_E(\sigma_1|q_2) \\ P_E(\sigma_2|q_1) & P_E(\sigma_2|q_2) \end{bmatrix}$$

Together, the states, emission probabilities, and transition probabilities define the model:

$$M = \{Q, P_E, P_T\}$$

$$\alpha(q,t) = P_E(\sigma_t|q) \sum_{\dot{q} \in Q} [P_T(q|\dot{q})\alpha(\dot{q}, t-1)]$$

Equation 1.3: The update rule at the core of the forward algorithm. The acoustic input at time t is $sigma_t$. □

For the example of figure 1.9, the update rule in matrix-vector notation is

$$\begin{bmatrix} \alpha(q_1,t) \\ \alpha(q_2,t) \end{bmatrix} = \begin{bmatrix} P_E(\sigma_t|q_1) & P_E(\sigma_t|q_2) \end{bmatrix} \begin{bmatrix} P_E(\sigma_1|q_1) & 0 \\ P_E(\sigma_2|q_1) & 1 \end{bmatrix} \begin{bmatrix} \alpha(q_1,t-1) \\ \alpha(q_2,t-1) \end{bmatrix}$$

Composite Model for Continuous Digit Sequences

For continuous speech, models are combined in a loopback arrangement (figure 1.10). This configuration allows a transition from the final state of each model into the start state of any other model.

Viterbi Decoding for Continuous Speech Recognition

$$\gamma(q,t) = P_E(\sigma_t|q) \max [P_T(q|\dot{q})\gamma(\dot{q}, t-1)]$$

Equation 1.4: The update rule at the core of Viterbi decoding; the forward algorithm's sum is replaced by a max operation. □

Parameter Estimation for HMMs

Hidden Markov models are generally trained using the expectation-maximization (EM) algorithm (Baum 1972) (Dempster et al. 1977), which was used here as well.

Continuous Emission Densities: CDHMMs

Continuous emission densities, used in all the HMM-based recognizers of this thesis, are not to be confused with continuous-speech recognition problems. The latter refers to the lack of pauses

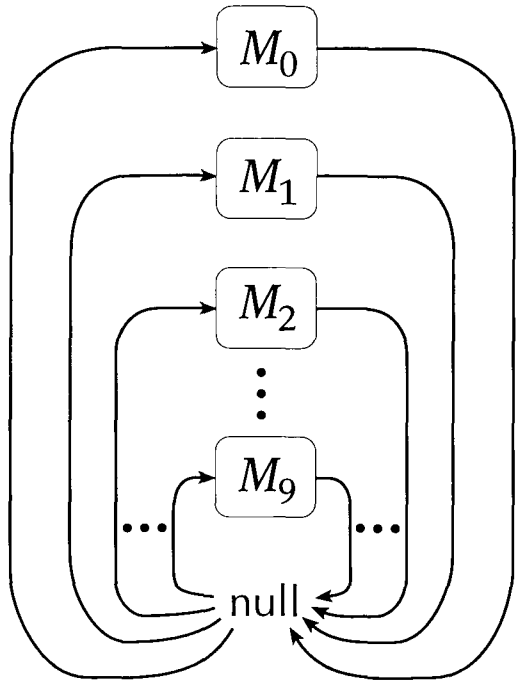


Figure 1.10: Composite model for continuous digit sequences: digit HMMs M_0 through M_9 are connected in a loopback arrangement. From the final state of every digit model, a transition is possible into the start state of any other model. □

between words in natural speech. The former refers to continuous parameters passed between the front end and back end—implemented in practice with floating-point numbers. Typically there is a vector of continuous parameters for each acoustic-analysis frame.

When each input frame is a vector instead of an integer—when the vector quantization step is skipped—the emission probabilities P_E defined above are replaced by a set of emission PDFs p_E. Each PDF typically (and in this thesis) is modeled as a weighted sum of Gaussians.

$$p_E(\mathbf{o}|q) = \sum_{i=1}^{G} w_{(q,i)} \frac{1}{\sqrt{(2\pi)^D \left|\mathbf{C}_{(q,i)}\right|}} e^{-\frac{1}{2}(\mathbf{o}-\mu_{(q,i)})^T (\mathbf{C}_{(q,i)})^{-1} (\mathbf{o}-\mu_{(q,i)})}$$

Equation 1.5: Mixture-of-Gaussians emission PDF for CDHMM. Each state q includes G Gaussians in its mixture. The means μ, covariance matrices $\mathbf{C}$, and mixture weights w are unique to each Gaussian. Each input vector $\mathbf{o}$ has D dimensions □

The weights are chosen such that

$$\sum_{i=1}^{G} w_{(q,i)} = 1$$

$$w_{(q,i)} \geq 0$$

1.11.3 Engineering Units of Speech: Monophones and Triphones

Most large-vocabulary continuous speech recognizers use a triphone representation of speech. A triphone $L\sigma R$ is a context-dependent unit, representing a phone σ when the preceding sound belongs to the set L and the following sound belongs to the set R. An acoustic model is trained for each triphone defined in the recognizer. In theory, L and R could each consist of a single phone. Such a complete enumeration would result in too many models for the available data, so in practice they are sets defined by automatic clustering during training (Jelinek 1998b).

In automatic speech recognition, the term *phone-like unit* is sometimes used instead of *phone*. The distinction is that it may be expedient to mix phonetic and phonemic categories when the goal is engineering rather than linguistics. Not every allophone needs to be enumerated. Still, the engineering units tend to be closer to a phone level than a phoneme level.

Monophone modeling, as used in chapter 3, refers to training models for each phone-like unit without regard to the neighboring sounds. It is considered a context-independent representation.

1.11.4 Grammar Modeling

Automatic dictation systems (large-vocabulary continuous speech recognizers) make use of simple top-down grammar models—based on bigrams and trigrams—during recognition. The recognizers created in this this do not use such models, but use unigram estimates of word frequencies (section 3.9.1). Top-down models are commonly called "language models" by researchers, but the present thesis avoids this terminology because bottom-up processing is also a part of language. This thesis proposes no new techniques for modeling at the grammar level; articulatory representations and lipreading fit naturally into the bottom-up side of recognition.

The state of the art in grammar modeling for recognition purposes is to estimate word-to-word (bigram; figure 1.11) and word-pair-to-word (trigram; figure 1.12) transition probabilities (Jelinek 1998b). The approach certainly does not capture all the grammar-level structure of natural language, but it has proven useful in the context of recognition.

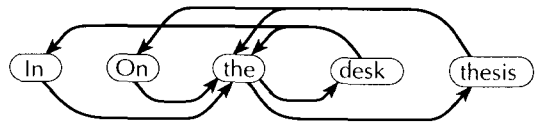


Figure 1.11: Bigram state machine—simplified example. This is a special case of a Markov model figure 1.8 in which transition probabilities are either uniform (which appear in this figure) or 0. □

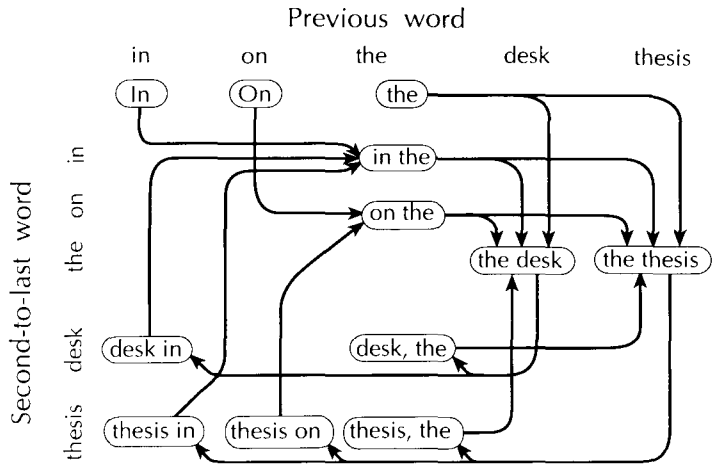


Figure 1.12: Trigram state machine. In a trigram Markov model (the simplified grammar generally used for speech recognition), all 30 word pairs would be allowed, and all 900 possible transitions. The transitions depicted here would have the highest probabilities. □

1.12 Historical Notes

1.12.1 Origins of Automatic Speech Recognition

In the nineteenth century, Alexander Graham Bell tried to solve the problem of enabling the deaf to understand speech sounds; a speech-to-text recognizer would have been ideal for the purpose, but with the technology of the time he could only build devices that presented the speech waveform visually. Such devices were known as phonautographs, and unlike the phonograph, could not play back sounds. Bell's phonautograph was unique in that the bristles that traced the waveform were connected to the bones of a human ear, which had been taken from a cadaver.

The first machine to recognize and selectively respond to speech was probably the Radio Rex toy dog (Paget 1930), manufactured by the Elmwood Button Company circa 1918-1929. It was designed to move when called—a binary decision. Improvements in signal processing made it possible

for automatic recognizers to distinguish multiple words (Davis et al. 1952). Increased computer performance enabled statistical models (Jelinek 1998b) to displace template-matching methods (Sankoff and Kruskal 1983), and recognition of continuous speech instead of isolated words.

1.12.2 Articulatory Symbols for Writing

Korean *han'gŭl* writing depicts articulatory characteristics of consonants in their written shapes (Kim-Renaud 1997). It was introduced by King Sejong of Korea in 1446 as a way to write Korean (and Chinese, for which it never gained popularity) without Chinese pictographic symbols. The stated purpose of the alphabet was to make writing easier and increase the literacy rate. Each consonant's letter form indicated which articulator was critical for producing the sound. For example, bilabial consonants were variations of a box shape, indicating the two lips, and dental consonants were based on an inverted V, indicating the bottom teeth. Each syllable was written by grouping vowel, consonant, and coarticulation symbols together in a box.

Linguists have devised schemes for transcribing speech in terms of articulation (Pike 1943)—for example, place of and type of constriction, and organ involved (Jespersen 1914). Alexander Melville Bell's Visible Speech (figure 1.13), like the Korean alphabet, attempts to graphically represent the articulator configuration (Bell 1867) for each character.

1.12.3 History of Speech Synthesis

There is a long history of speech synthesis, both acoustic (von Kempelen 1791) and electrical (Stewart 1922). The acoustic synthesizer was preceded, in European music, by the vox humana stop (setting) of the organ. The latter produces an ethereal sound that resembles a voice or a chorus singing a single, sustained vowel sound; while the goal of synthesis is to produce intelligible speech.

Synthesis has, from the start, been based on a source-filter model of speech. The filter represents the resonances of the vocal tract, and the source (in voiced sounds) represents the vibration of the vocal cords (figure 1.14). Initially the filter was implemented in an acoustic resonator designed to have a single formant frequency (von Kempelen 1791). Later synthesizers used pairs of formants, either for a small subset of sounds (Stewart 1922) or for all sounds except /h/ (Paget 1922). In the twentieth century, electrical synthesis largely replaced acoustic, and then digital electronic synthesizers succeeded analog. A milestone of synthesis was the demonstration of an electrical synthesizer, with a human operator controlling analog parameters in real time, at the World's Fair (Dudley et al. 1939).

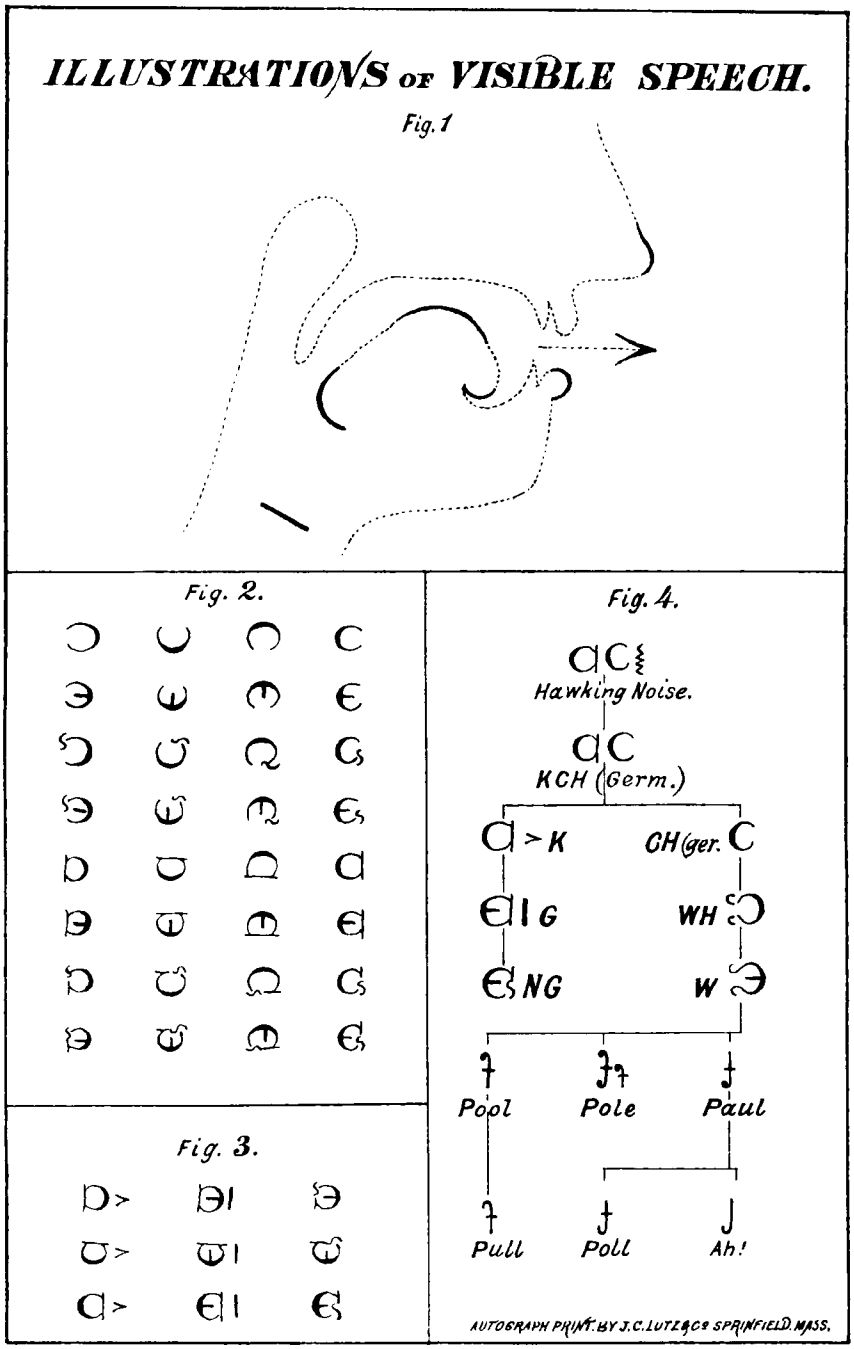


Figure 1.13: Visible Speech, an alphabet representing articulations graphically (Bell 1867); one of several alphabets designed on this principle to facilitate learning; predated by Korean *han'gŭl* alphabet (1446). Image digitally retouched from a U.S. Library of Congress scan. □

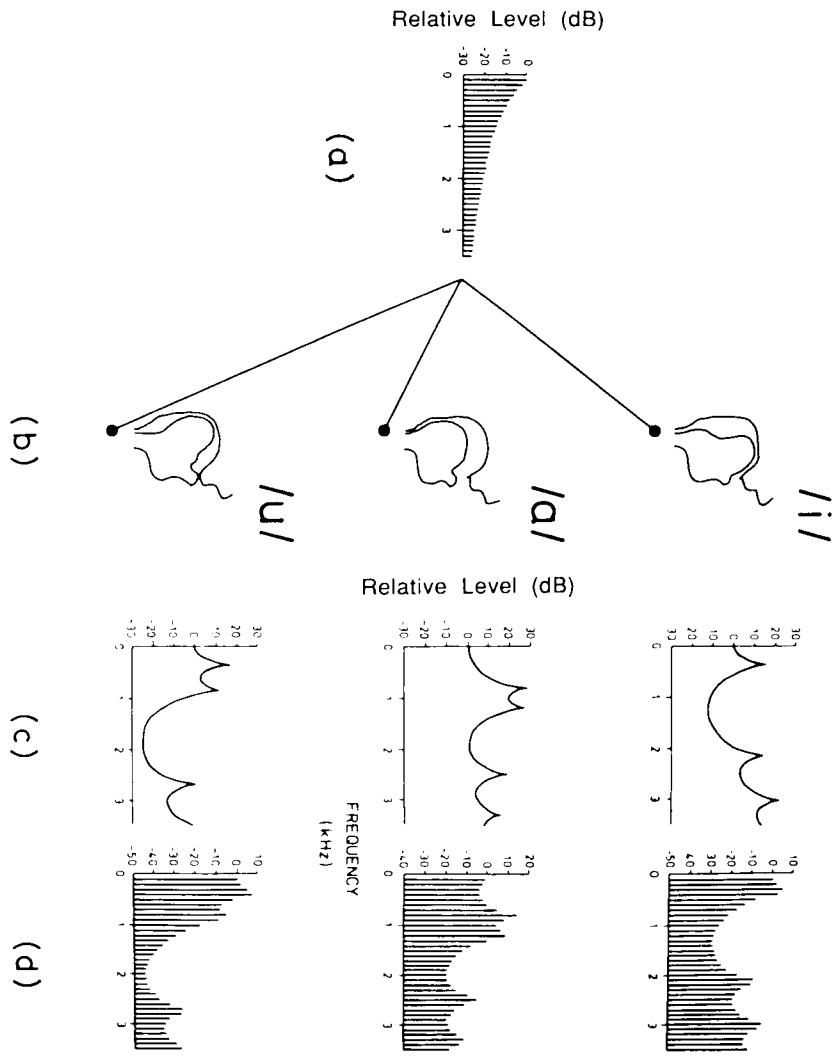


Figure 1.14: Source-filter model of speech synthesis, used in various forms for hundreds of years (von Kempelen 1791). The vocal cords in isolation would produce a sound with a spectrum like (a) in the figure; different mouth shapes associated with the various phonemes (b) cause resonances as depicted in (c). The linear filtering of (a) by (c) gives speech signals whose spectra look like (d). Reprinted with permission (Bailey 1983); copyright 1983, Academic Press. □

Contemporary synthesizers use digital signal processing techniques (Breen 1992).

1.13 Third-Party Assistance

The author was the original grant author and principal investigator for the lipreading work (chap. 2), which was performed at Tanner Research, Inc. (Pasadena, CA, USA) under U.S. Air Force Small Business Innovation Research (SBIR) Contract F41624-97-C-6017 (Fain 1998). A patent is pending with Tanner Research as the assignee (Fain and Chinn 1999). The proposal described a customized face mask, high-contrast imaging, thresholding, edge identification, and centroid computation. Garry Chinn suggested some changes to the image processing pipeline: specifically, the median filter for noise, and the technique to find the corner of the mouth; he also did a considerable amount of implementation. Jayant Shukla modified Unix shell scripts for acoustic recognition; he also collected a second data set, results for which are not reported here, except for the raw waveforms pictured in figure 2.11.

The author applied optimal feature extraction (appendix B) to large-vocabulary continuous speech recognition at Johns Hopkins University (Baltimore, MD, USA) as part of the 1997 Summer Workshop of the Center for Speech and Language Processing. Terri Kamm helped modify the baseline HMM training procedure to use the resulting feature transformation. The technique was proposed by Nagendra Kumar (Kumar and Andreou 1998).

1.14 Summary of Results of Thesis

The implications of the results of this thesis are discussed in chapter 5. One result presented here is that continuous speech recognition can be improved by augmenting acoustic recordings with direct articulator measurement (chap. 3). Also reported is a new device and processing pipeline for side-view lipreading (chap. 2); the device and algorithms were designed together for robust feature extraction. The two types of kinematic data—direct measurement and extraction from video—both represent up-down and front-back motion of articulators. Lipreading fits naturally in an articulatory framework because it allows motion to be inferred, and because more-direct measurement is very invasive (appendix E). The present experiments did not involve inferring motor representations from sound, but other work has demonstrated the feasibility of recovering such information.

Kinematic information was found to improve vowel recognition more than consonant recognition,

although vowels are often thought to be better modeled by existing recognizers (chap. 4). This result inspired a reexamination of the performance of some standard state-of-the-art recognizers, which revealed no advantage in recognizing vowels over consonants. The relative information content of vowels and consonants in text was calculated, and stops and fricatives together were found to convey about as much information as vowels. The information-content results are somewhat at odds with the common assumption about vowels and consonants; that assumption depends on including not only stops and fricatives, but also vowel-like sounds, in the category of consonant.

A front-end optimization technique, heteroscedastic discriminant analysis, was found to scale from a smaller problem (continuous digit recognition), where it had been previously demonstrated, to transcription of telephone conversation. Theoretical difficulties were discovered in applying the technique to HMMs having mixture-of-Gaussian emission PDFs.

Chapter 2

Side-View Lipreading Device and Algorithms

2.1 Problem Description

The goal of this chapter's project has been automatic lipreading: speech recognition from audio and side-view video (Fain 1997) (Fain and Chinn 1999). In order to simplify image processing—to make it more robust and less computationally demanding—a custom side-view input device was developed to acquire video of the mouth's motion. A new video processing pipeline is presented that segments the upper and lower lip regions and computes their centroids, which in turn are useful for recognition (see confusion matrix in table 2.2).

Recognizers for the audio and video data were developed, and errors were analyzed across modalities (similar to the separate-identification technique for merging audio and video). The recognizer's errors in the two modalities were completely disjoint: the two categories confused by the video pathway were never confused in recognition from audio alone. This indicated that for a very small data set, audio and side-view video provided complementary information.

The input device is a pilot's oxygen mask modified to include a miniature video camera and light. The approach is directly applicable to situations in which the user is required or willing to wear an

opaque mask (e.g., fire fighting, surgery, or piloting an unpressurized airplane). It might be possible to generalize the technique to a translucent mask or an open headset.

2.2 Background

2.2.1 Degrees of Freedom of Lip and Jaw Motion

In designing a lipreader, it is worth considering how many independent degrees of freedom (DF) the lip and jaw have; this chapter uses 4 DF. The idea that there might be a tractable number of DF comes form the fact that there are a limited number of muscles in the face. However, motions may involve the coordinated action of several muscles, and some muscles may move in more than one way. The widely used facial action coding system identifies 19 possible movements of the lip and jaw (appendix D). Many of these motions are emotional expressions that are not required to make phonetic distinctions, so the number of effective DF related to speech is probably fewer.

For front-view lipreading, the tongue is certainly relevant, since it can sometimes be seen when the mouth is open. This project's side views make the tongue harder to see. Raising and lowering the tongue to make different sounds is correlated to jaw motion, so even from the side it may be possible to extract some information about the tongue's motion.

As described in section 2.5.6, the present project transforms the side-view video into four parameters per frame. These parameters are two pairs of coordinates, representing the center positions of regions around and including the upper and lower lip.

Other researchers have attempted to determine how many degrees of freedom are sufficient for lipreading by humans (Benoît et al. 1996). They presented listeners with speech in noise, either alone, with animation, or with video of the actual speaker. Intelligibility tests were repeated for animation of the lips alone, the lips superimposed on a skull with a moving jaw, or an entire face. The face and skull-plus-jaw models had the same 6 DF of motion; the lip model had only five. These motions were measured reliably from the speaker's face, which had blue makeup applied on the lips and three landmarks. For a 0 dB signal-to-noise ratio (SNR), the 6 DF models were as helpful for lipreading as the actual face. At –18 dB SNR, speech was essentially incomprehensible without video, and seeing the human (not synthetic) face reduced errors by 25% relative to the 6 DF face animation. This discrepancy may be caused by the unnatural appearance of the animated face, or the specific choice of parameters.

Video		Error rate by Audio noise level	
Type	DF	Quiet (0 dB SNR)	Noisy (−18 dB SNR)
None		39.74%	98.68%
Animated lips	5	27.81%	77.48%
Animated skull, lips	6	21.85%	70.20%
Animated face	6	20.53%	54.97%
True speaker		23.18%	41.06%

Table 2.1: Previous work (Benoît et al. 1996): lipreading error rates by human listeners for different face parameterizations by human listeners. Faces were animated with either 5 or 6 degrees of freedom (DF), and speech with 0 or −18 dB signal-to-noise ratio (SNR) was played back. At the latter SNR, comprehension from audio alone is almost impossible. The 6 DF model improves recognition considerably; only 4 parameters represent lip configuration in this chapter. □

2.2.2 Distinctive Phonetic Features Relating to Lips and Jaw

Another way to estimate the degrees of freedom of the lip and jaw is to approach the problem phonetically, rather than physiologically. The tongue conveys a tremendous amount of phonetic information and its position may be inferred to some extent even from a side view, because tongue motion is correlated to jaw motion.

2.3 Previous Work

2.3.1 History of Machine Lipreading

Forty years ago, a patent entitled *Electronic Lip Reader* was filed (Nassimbene 1965). That system, like the one described in the present thesis, was to be worn on the head. The device was intended for use with speech recognition, although the patent did not specify a classification technique.

Lipreading has made considerable progress since then, with many projects concentrating on recognition from video under general lighting conditions and variable camera position (Hennecke et al. 1996).

Previous work helps illustrate the link between this chapter and chapter 3 (Westbury and

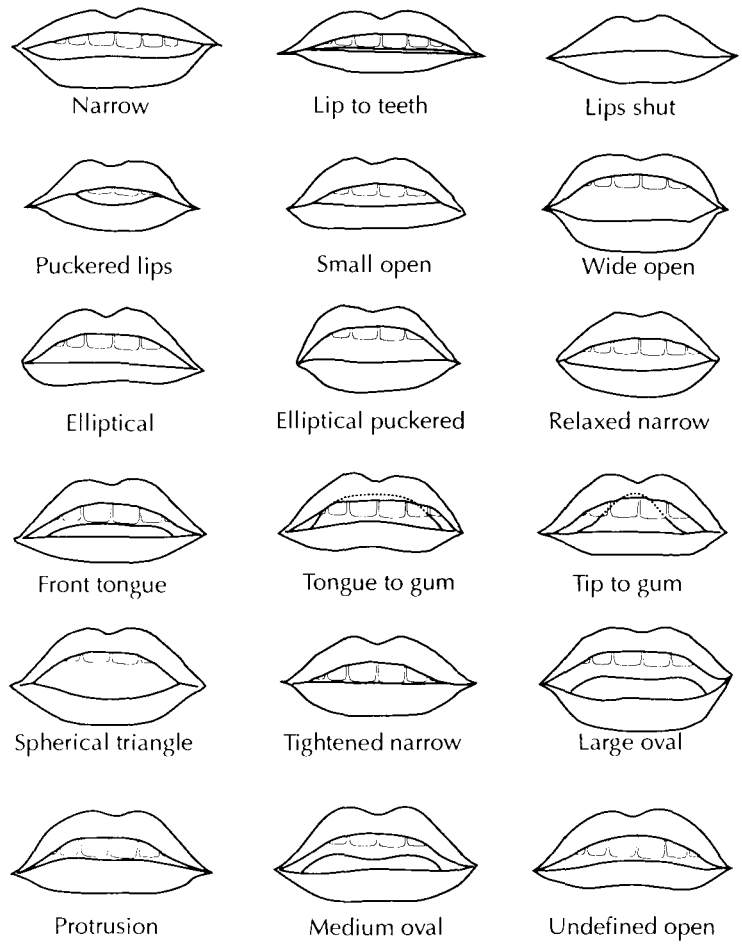


Figure 2.1: Lip shapes associated with different sounds of speech, as seen from the front as opposed to the side view of this thesis. Reprinted with permission (Parke and Waters 1996); copyright 1996, A. K. Peters. □

Hashi 1997). In that project, lip motion was found to be similar between English and Japanese speakers making comparable consonant sounds, and to convey considerable phonetic information. It used the same database of x-ray microbeam recordings of metal beads on the lips as used in chapter 3.

2.3.2 Different Ways to Merge Audio and Video

Different ways to merge audio with video for lipreading are defined in section 1.7: direct identification, separate identification, motor-space recoding, and dominant recoding (Robert-Ribes et al. 1996). The present chapter uses the separate identification approach. In previous work, sepa-

rate identification has been found to outperform direct identification (Adjoudani and Benoît 1996) (Massaro 1996), but motor-space recoding also shows promise (Robert-Ribes et al. 1996). Lipreading by humans is described in appendix C.

2.3.3 Dynamic Contours (Snakes)

Much recent work in lipreading has used dynamic contours, also known as snakes (Hennecke et al. 1996). A snake tracks a curved contour as it moves in a video input (Kass et al. 1987). After a snake is initialized to lie on an edge in an image, its subsequent motion is defined by energy-minimizing differential equations, derived via the calculus of variations. When constraints apply to the contours—for example, the symmetry of the two edges of a finger—the equations of motion can be augmented so that the constraints will be asymptotically satisfied (Platt 1989).

2.3.4 Camera Placement

Although most previous work has considered the more general case of camera placed order 1 meter away from the subject (Hennecke et al. 1996), some other researchers have used wearable cameras to control the recognition problem (section 2.3.1) (Bass et al. 1997).

A wearable, video-only lipreader was mentioned in passing in the summary of a research workshop (Bass et al. 1997); however, neither literature nor Web searches revealed any further information on the system.

2.4 Lipreading Face Mask

In this chapter's project, the user wears an opaque mask, which includes a miniature camera, light, and noise-canceling microphone. Mounting the camera and light to the mask increased the total weight from 372 g to 394 g. The helmet kept the mask in place, which could alternately be accomplished by a strap.

The camera and light are both placed to the side of the user's mouth. This results in a side-view image, instead of the front view common in lipreading research.

Video data from the camera are passed through several sequential stages of processing. The initial front-end stages estimate the positions of regions of the face. The back end then classifies speech based on this position information.

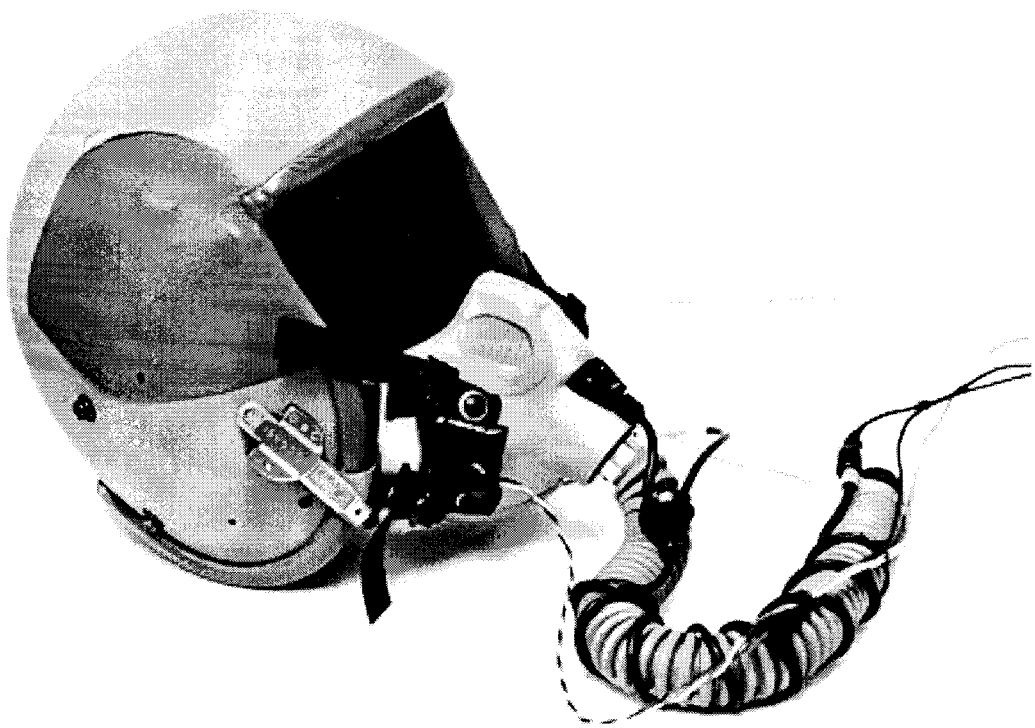


Figure 2.2: Side-view lipreading face mask designed and built in this project. The light is inside the mask, and the video camera is on the mask's right side, covered in electrical tape. Three twisted wires lead away from the camera. The helmet keeps the mask in place. Image copyright Tanner Research; used with permission. □

2.4.1 Design Issues

The requirements for the face mask design were:

1. Facilitation of simple, robust image processing—thus, high-contrast, side-view video acquisition
2. Light weight
3. Low profile—low polar moment of inertia with regard to the neck axis

2.4.2 Alternate Designs Considered: Silhouette Imaging

Other designs considered for the lipreading face mask were using an array or single point sensor (nonfocusing), or illuminating the mouth on the opposite side from the image sensor, producing a back-lighted silhouette (Fain 1997). In order to detect the shadow, either the light source or the sensor would have to be extended spatially. Some possible silhouette imager configurations would have been:

- A diffuse illuminating panel with a video camera.
- A set of light sources with low switching times (such as light-emitting diodes) lit in a repeating sequence, with a point sensor determining whether it was in shadow at a given time.
- A spread-out array of sensors used in conjunction with a switching array could capture three-dimensional information—the shadows produced on the opposite side of the mask from the switching array would show a parallax effect, which could be used to reconstruct shape.

Ultimately, the easy availability of a small video camera—and the simplicity of using a single light bulb instead of a diffuse illuminator or an array—motivated placing the light on the same side as the camera. Even the parallax-sensing configuration mentioned above is arguably no better than simultaneous readout from multiple cameras.

2.4.3 Implementation

The lipreading face mask was constructed by cutting a hole in the side of the oxygen mask, inserting the lens of a compact video camera through the hole, and attaching a light to the inside of the face mask on the same side as the camera. The video camera was a commercially available part; its

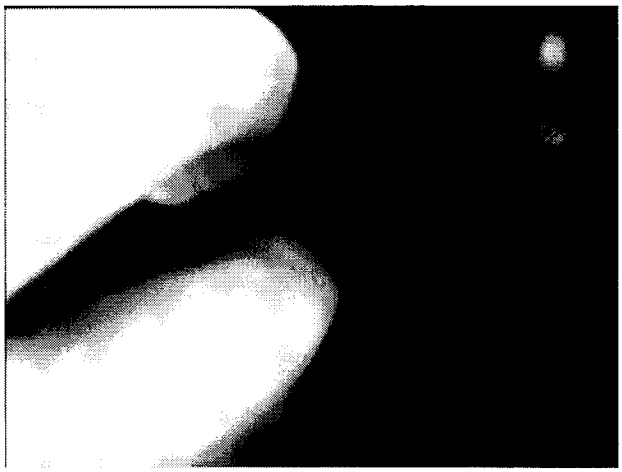


Figure 2.3: Still image of subject saying /i/ ("ee"). This and other images collected as part of this chapter's project are copyright Tanner Research and used with permission. □

circuitry fit on a single, small circuit board (figure 2.2) and it produced National Television Standards Committee (NTSC) output. The light was a small incandescent bulb.

2.5 Front-End Feature Extraction Using Face Mask

For this project's front end, a desktop computer acquires the camera's video signal using a video-capture expansion board, which samples 30 frames per second at 320 (horizontal) by 240 (vertical) resolution. Examples of the video recorded by the camera appear in figure 2.3 through figure 2.6, for four different phonemes. The distinction of /i/ (figure 2.3) versus /u/ (figure 2.3) was chosen because it involves lip rounding and could hypothetically pose a problem for a side-view recognizer such as the one of this chapter. The labiodental sound /f/ (figure 2.6), on the other hand, involves a distinctive lip-tooth configuration.

The front end developed in this project includes the following processing steps:

- Noise removal;
- Threshold to separate face from background;
- Segment upper and lower regions;
- Compute centroids for both regions.

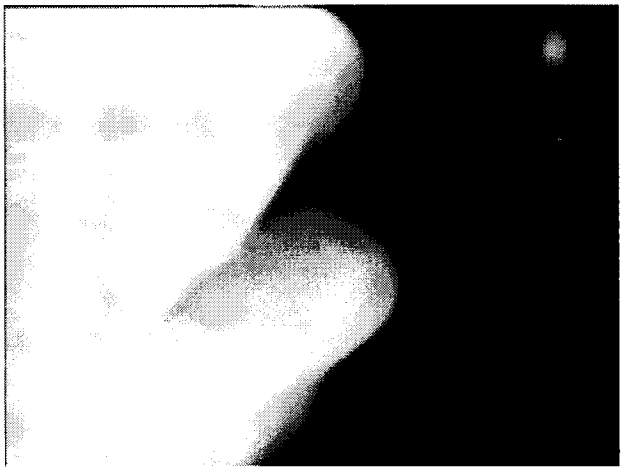


Figure 2.4: Still image of subject saying /u/ ("oo"). Compared to figure 2.3, the lips have moved together and forward. The use of the side view to distinguish these phonemes appears in figure 2.10. □

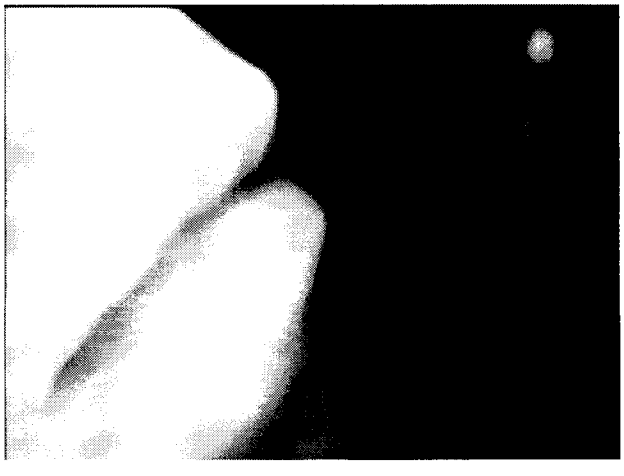


Figure 2.5: Still image of subject saying /m/. As expected, the lips are closed against each other. □

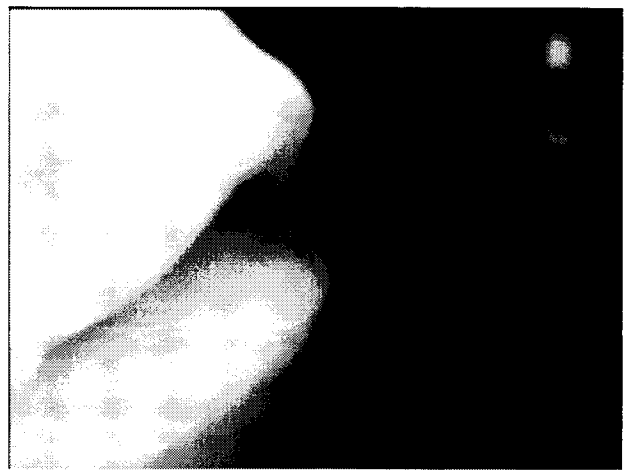


Figure 2.6: Still image of subject saying /f/. The lower lip has moved back to touch the bottom of the upper teeth. □

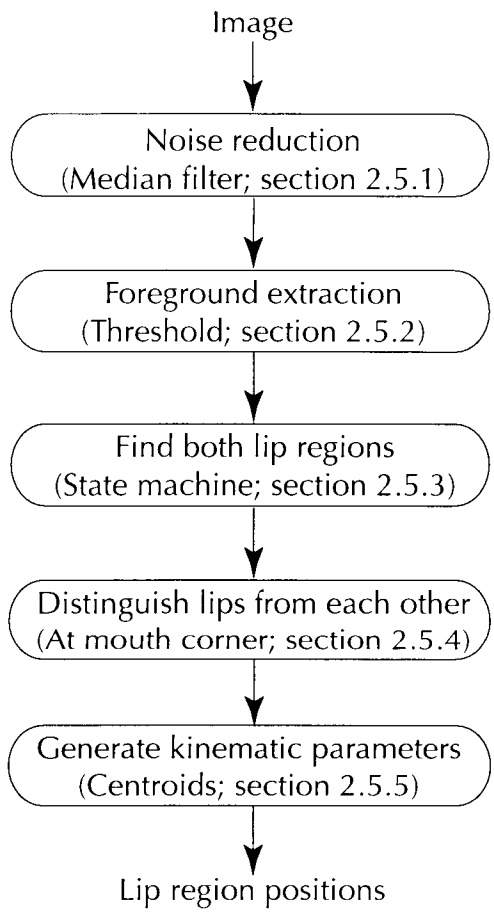


Figure 2.7: Video-processing steps for the lipreading experiments described in this chapter. □

The processing occurs independently on each video frame, so in the following discussion, a frame will be represented by $I(x, y)$, which is defined for integers x and y in the field of view.

2.5.1 Noise Removal with Median Filter

The video acquired by the computer may contain speckled noise artifacts. To eliminate these, two median filters, one horizontal and one vertical, are applied in sequence. Each filter operates on three-pixel intervals. The center pixel of each interval is replaced by the median-valued pixel in the filter window.

For the horizontal filter operating at (x, y),

$$\alpha \doteq I(x-1, y)$$
$$\beta \doteq I(x, y)$$
$$\gamma \doteq I(x+1, y)$$

Similarly, for the vertical filter,

$$\alpha \doteq I(x, y-1)$$
$$\beta \doteq I(x, y)$$
$$\gamma \doteq I(x, y+1)$$

The output of median filtering is as follows:

$$I_2(x, y) = \alpha \text{ if } \beta \leq \alpha \leq \gamma \text{ or } \gamma \leq \alpha \leq \beta$$
$$I_2(x, y) = \beta \text{ if } \alpha \leq \beta \leq \gamma \text{ or } \gamma \leq \beta \leq \alpha$$
$$I_2(x, y) = \gamma \text{ if } \alpha \leq \gamma \leq \beta \text{ or } \beta \leq \gamma \leq \alpha$$

2.5.2 Thresholding Face from Background

The prototype mask has produced high-contrast images of the speaker, so a threshold has been sufficient to highlight the speaker's face against the background, while maintaining the border between the lips with the mouth closed (figure 2.5). The a priori threshold of one-half the maximum intensity worked well, resulting in a binary image for the next stage of processing.

$$\theta = \frac{\max I_2(x, y)}{2}$$

$$I_3(x, y) = 0 \text{ if } I_2(x, y) < \theta$$
$$I_3(x, y) = 1 \text{ if } I_2(x, y) \geq \theta$$

2.5.3 State Machine Distinguishing Upper and Lower Lips

A state machine was used to determine an upper and lower bound for each of the lip regions. These bounds give overlapping regions that are converted to distinct regions as described in section 2.5.4.

The state machine is defined in figure 2.8. The boolean state variable a indicates whether the machine is currently (i.e., for a particular position x, y) inside a lip region; b is true if the state machine is at or past the the end of the region. The start (bottom edge) of the lower lip region, for a particular horizontal position x is $S_L(x)$; the start (top) of the upper lip region is S_U. The ends of the two regions are E_L and E_U. Therefore, the contour of the overlapping lower-lip region is the following set of points:

$$\{(x, S_L(x))\} \cup \{(x, E_L(x))\}$$

2.5.4 Finding Corner where Lips Meet

The corner where the two lips meet was defined as $(x, E_L(x))$, in which x was the farthest-left horizontal position for which the two overlapping regions did not overlap. This was very close to $((x), E_U(x))$, as indicated in figure 2.9.

2.5.5 Calculating Centroids of Lip Regions

The final step of front-end processing is to compute the centroids (center of area) of the upper and lower regions. For the binary, sampled image representation of each lip's region, the centroid is computed as follows:

$$x_C = \sum_{y=1}^{h} \sum_{x=1}^{w} b(x, y)x$$
$$y_C = \sum_{y=1}^{h} \sum_{x=1}^{w} b(x, y)y$$

Equation 2.1: Centroid coordinates (x_C, y_C) of a sampled, $h \times w$ binary image b □

The video processor's output is two coordinate pairs—(x_U, y_U) and (x_L, y_L)—the first pair obtained by substituting I_5U for b in equation 2.1, and the second by substituting I_5L.

```
for x := 1...w
    if I3(x, 1) = 0
        q := 0
    else
        q := 1
        SL(x) := 0
    b := 0
    for y := 1...h
        if b := 0
            if q := 0 and I3(x, y) = 1
                q := 1
                SL(x) := y
            if q := 1 and I3(x, y) = 1
                q := 0
                b := 1
            EL(x) := y
```

Figure 2.8: Pseudocode for segmentation of lower lip region. For upper lip, direction of inner **for** loop is reversed, from $h...1$; and S_U and E_U are substituted for S_L and E_L. □

2.5.6 Features Generated by Front End

Output of the front-end image processor is shown in figure 2.10 for a single frame from each of multiple experimental trials. In this case, the user repeatedly spoke the four sounds /m/, /f/, /i/ ("ee"), and /u/ ("ooh" as in "shoe"). The graph is oriented to represent a speaker facing right, and only the lower region's centroid is depicted.

2.5.7 Robustness to Mask Placement

The subject removed and replaced the mask repeatedly during data collection. Any inconsistency in position between the different mask placements was too small to interfere with classification. This

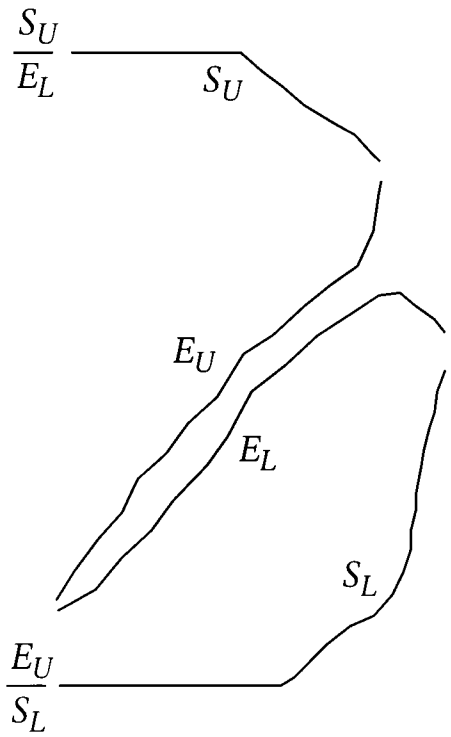


Figure 2.9: Segmentation of lower, upper lips, as determined by state machine of figure 2.8. In the left part of the image the two regions are overlapping; the final step is to split them (see text). For illustration purposes, gaps have been drawn between contours which are immediately adjacent to each other. □

may be due to the fit of the mask to the face, or to the distinctiveness of the categories when measured with this feature set, or, most likely, both. Harder recognition tasks may require algorithmic calibration and different mask styles.

2.6 Centroid Motion Versus Lip Motion

Movement of the lower centroid is related but not identical to movement of the lower lip. The centroid moves as predicted by phonetics. Compared to /m/, in which the lips are closed, the labiodental /f/ is produced with the region shifted down and back, /i/ with the region shifted down, and /u/ with the region shifted forward (figure 2.10).

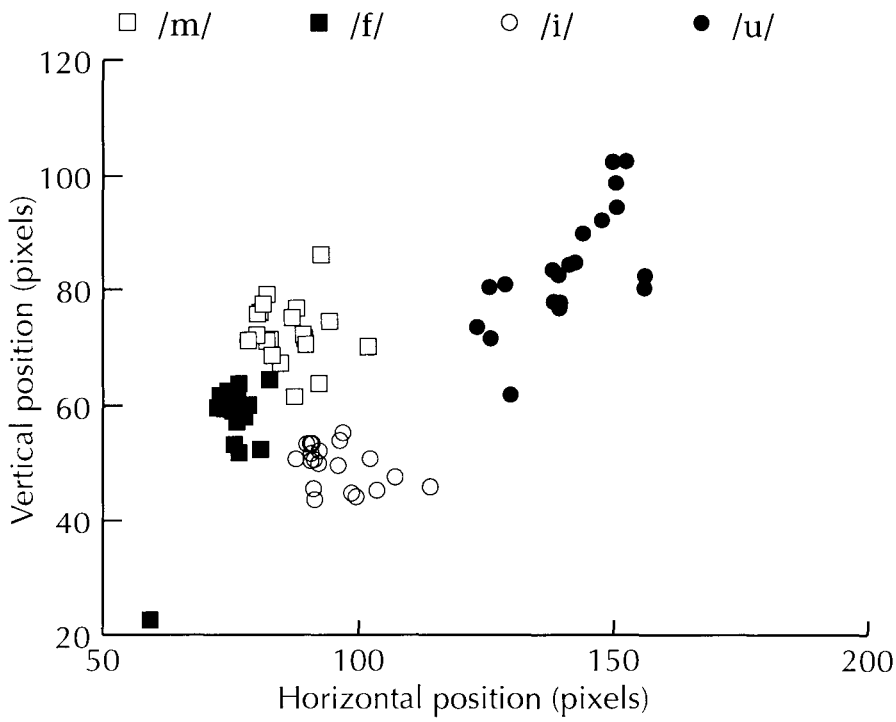


Figure 2.10: Position of centroid of the lower-lip region as the subject (facing right) makes the four sounds /m/, /f/, /i/, and /u/. See table 2.2 for classification results for these data. □

2.6.1 Aperture Effect

Due to an aperture effect, the rate and direction of motion of the centroid will differ from the motion of the lips themselves, and from a landmark on the lips such as used in chapter 3. This effect can be illustrated with a couple of simple examples. It is not representative of the aperture effects seen in human perception; it is due to the combination of a finite image frame and the specific centroid calculation used for this project. Although the lips have the added complication of being nonrigid, the aperture effect is seen for rigid (i.e., nondeforming) shapes such as these examples.

Aperture Effect on Rectangle

Consider a rectangle that is only partially in an image frame to begin with, that is oriented parallel to the edges of the frame, and that starts moving parallel to an edge and out of the frame (figure 2.12). In this simple case, since the centroid is computed using only the visible portion of the rectangle, it will move at half the rate of the whole rectangle.

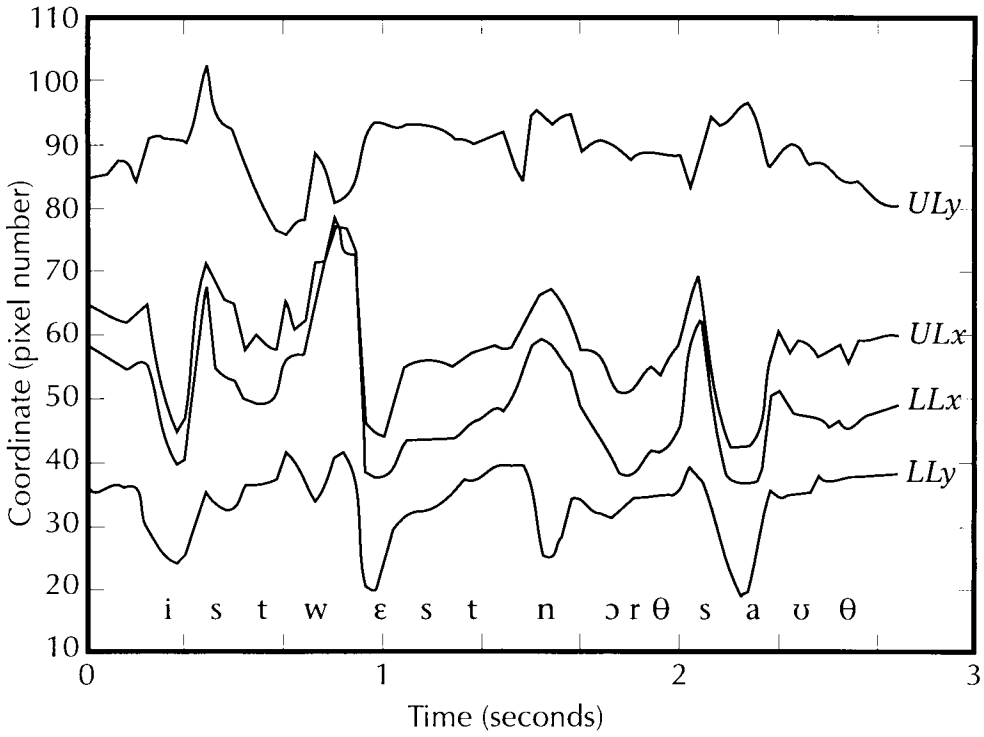


Figure 2.11: Changes in lip region coordinates over time, as a speaker says "east, west, north, south." Lower lip coordinates are (*LLx*, *LLy*), and upper lip coordinates are (*ULx*, *ULy*). □

Aperture Effect on Parallelogram

Not only the centroid's speed, but also its direction of motion differ from the same quantities measured at a landmark on the lips (such as the metal beads of chapter 3). A parallelogram moving out of the image plane illustrates this effect (figure 2.13). In this case, the centroid's upward speed is greater than half the parallelogram's; the centroid also moves left, although the parallelogram's motion is strictly vertical.

2.6.2 Comparison to Previous Methods

The front-end image processing of this project has several advantages over dynamic contours (section 2.3.3):

- It requires considerably less computation;
- It acts on frames independently, which prevents errors from propagating from one frame to the next;

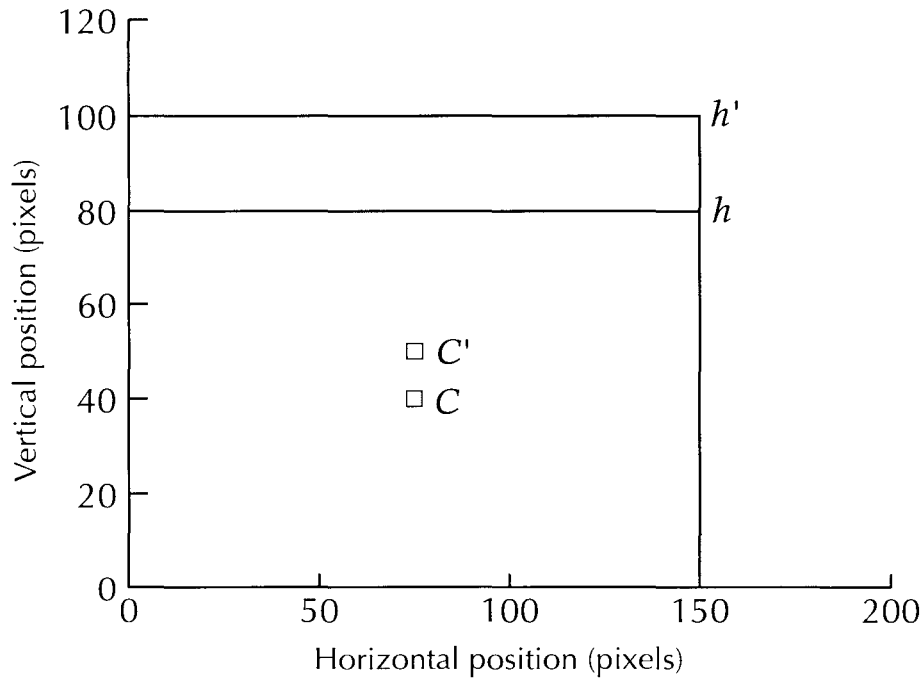


Figure 2.12: Aperture effect in centroid calculation with a rectangle moving upward from h to h'. The centroid moves from C to C' at half the rate of the rectangle, only part of which is visible. □

- No arbitrary parameter-setting experiments were required;
- There is no separate initialization.

One of the most closely related projects combined single-frame processing for the positions of the eyes and nose with frame-to-frame lip tracking (Petajan and Graf 1996). In that project, lip tracking was reset by single-frame calculations whenever the lips closed. Because frames were analyzed one at a time, and lip tracking was periodically reset, their video processor avoided the problems of poor initialization and indefinitely long propagation of tracking errors, and it performed well under a wide range of lighting conditions.

2.7 Simple Back-End Classifier

Two back-end classifiers have been implemented; one operates on single video frames, and the other on sequences of frames. However, problems with training data prevented evaluation of the frame-sequence recognizer.

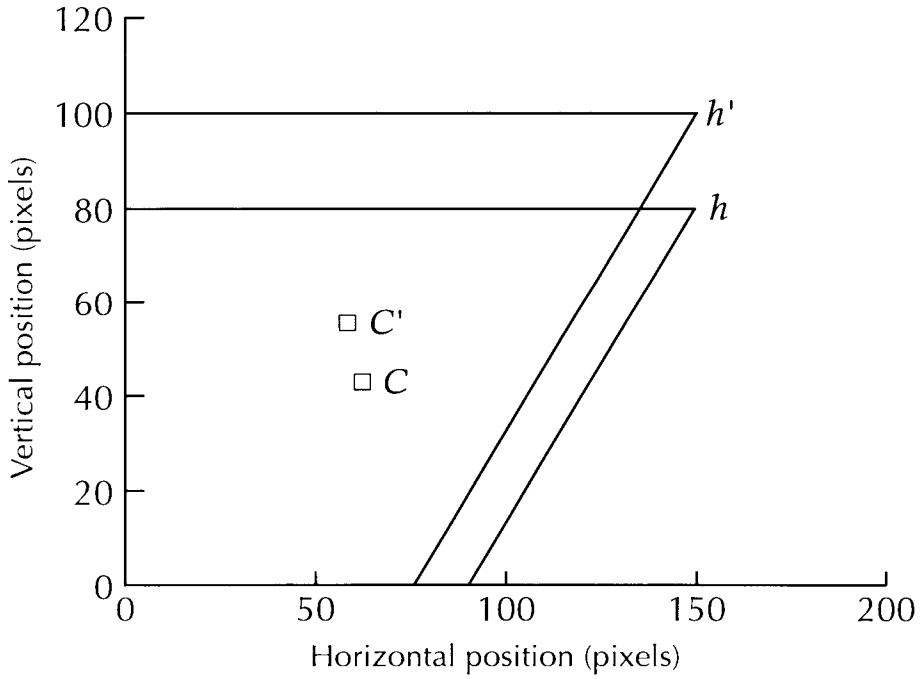


Figure 2.13: Aperture effect for a clipped parallelogram moving straight up from h to h'. The centroid moves left, and its upward rate is slightly faster than the rectangle's centroid in figure 2.12. □

2.7.1 Single-Frame Recognizer

A maximum-likelihood Gaussian classifier section A.1 (Duda and Hart 1973), with diagonal covariance matrices, was used to classify centroids from the front-end processor; recognition experiments and results are described in more detail below.

2.8 Lipreading Recognition Results

A pair of recognizers, one acoustic, one lipreading, were trained to determine whether the two modalities provided independent information about what was spoken.

Classification experiments were performed which quantified the separation between categories seen in figure 2.10; results of these experiments appear in table 2.2. For each example, maximum-likelihood Gaussian classifiers appendix Adefs were trained with that example omitted from the training set. Omitting the test example ensured a proper train/test split for evaluating generalization ability.

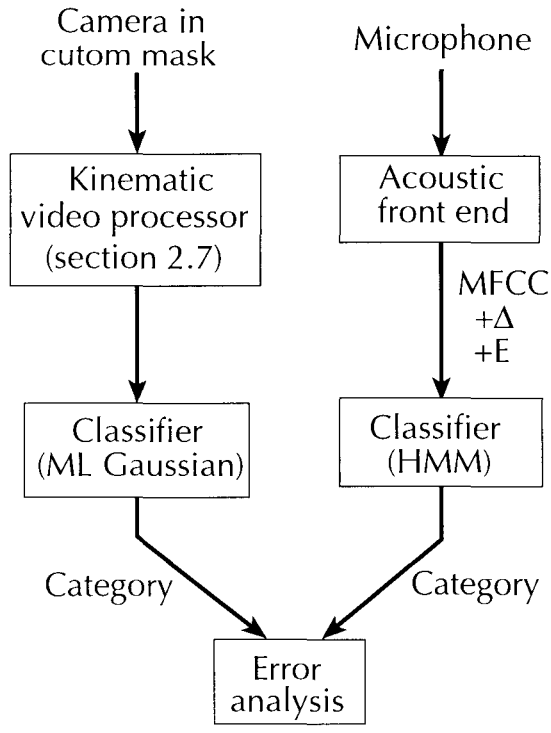


Figure 2.14: Lipreading architecture tested on a very small data set; results are reported in section 2.8.3. This architecture is an example of the separate identification approach (section 1.7.1), as opposed to the direct identification approach of chapter 3 or the motor-space recoding approach (Robert-Ribes et al. 1996). □

2.8.1 Acoustic Recognizer

Sound recordings were made of a single subject speaking four words: "north," "south," "east," "west." Twenty-five repetitions of each word formed the training set, and ten formed the test set. A conventional hidden Markov model acoustic recognizer was used, with continuous emission densities (CDHMM; section 1.11.2) and Mel-frequency cepstral coefficients (chap. 1) (Young et al. 1997).

Each CDHMM had three emitting states; each of those states had a mixture-of-Gaussians emission PDF with a diagonal covariance matrix. The front end features were the standard Mel-frequency cepstral coefficients with delta, acceleration, and energy terms (section 1.10.2). Training lasted for two iterations of the expectation-maximization algorithm (section 1.11.2).

The recognizer performed perfectly on the small data set, but in the target application, performance would be degraded by considerable noise, speaker-to-speaker variability, continuous speech, and a larger vocabulary. Noise was added to the acoustic data, and the recognizer was trained with

	Classifier output			
	/m/	/f/	/i/	/u/
True category: /m/	21	0	0	0
/f/	1	18	0	1
/i/	0	0	20	0
/u/	0	0	0	20

Table 2.2: Confusion matrix for classification of /m/, /f/, /i/, and /u/, using the video component only of the pipeline in figure 2.14. The error rate is 2%. See also figure 2.10. □

	Error rate
Audio in noise	17%
Video	10%
Combined	0%

Table 2.3: Error rates of audio and video recognition. Note that the joint error rate was not verified by cross-validation, but by error analysis. See also table 2.2 which reports a lower video-only error rate for a slightly different problem. □

data having 0 dB SNR and tested with data having 5 dB SNR. In these circumstances, the acoustic recognizer guessed wrong 17% of the time.

2.8.2 Still Image Recognizer

Video images were captured of the same subject making the initial sounds—/n/, /s/, /i/, and /w/—of the four test words. These images were used to train a maximum-likelihood Gaussian classifier (section 2.7).

Training used 20 cases each of /n/, /s/, /i/, and /w/. Five separate cases for each category were used for testing; the error rate on this test set was 10%.

2.8.3 Analysis of Errors by Modality

Comparing the specific errors made by the recognizers is more instructive than comparing the error rate. The video recognizer only confuses /s/ and /i/, while the acoustic recognizer never confuses "south" and "east." Thus the video recognizer can be used to decide whether a subject is speaking "north," "west," or something else, and then the audio recognizer can differentiate "east" and "south."

Although errors in this small problem could be eliminated completely, such results should not be expected for larger problems. The analysis simply reflects that, with the present data set, the two modalities provide independent information.

2.9 Side-View Lipreading: Conclusions

Above, preliminary recognition results and an error analysis show that combining audio and video in side-view lipreading improves performance. A camera and light positioned for a side view, coupled with simple processing of individual frames of video, result in a robust tracker of upper-lip and lower-lip position. Since almost all previous lipreading work has used a front view, the initial data included distinctions that might hypothetically be easier to see from the front: rounded/unrounded (/u/ versus /i/) and labiodental/bilabial (/f/ versus /m/). These distinctions remained clear in the parameters produced by this side-view lipreading front end; of the phonemes just mentioned, only /f/ was misclassified (table 2.2).

2.10 Future Work

The opaque mask is rather obtrusive for users who aren't required to wear a mask for other reasons, as, for example, fighter pilots are.

Further data collection, for additional speakers and more complex utterances, would enable a thorough assessment of this side-view, threshold-and-centroid approach to lipreading. Ideally, such a data set would include front-view video recordings as well, for comparison.

Chapter 3

Speech Recognition with Direct Articulatory Measurements

The goal of the experiments described in this chapter was to determine the benefit of adding articulatory information to a hidden Markov model (HMM) based continuous-speech recognizer. The motivation for adding an explicit kinematic representation is described in section 1.1.

The articulatory data used in these experiments were collected at the University of Wisconsin, by other researchers, using the x-ray microbeam technique described in section 3.4.1 and appendix E. This technique determines the back-to-front and bottom-to-top positions of points on the tongue, teeth, and lips. The Wisconsin data include a number of subjects reading words, sentences, and paragraphs while position information and sound are simultaneously recorded.

In the project of this chapter, concatenating transformed articulatory information to a standard acoustic (cepstral) representation reduced the error rate by 7.4%. This demonstrated across-speaker statistical significance ($p = 0.018$) for the first time in continuous recognition from articulatory data. The motion data improved recognition for male speakers more than female, and recognition of vowels more than fricatives or stops. The comparison between vowels and consonants is described in more detail in chapter 4, and the coordinate transform is analyzed further in appendix A.

Speaker-dependent monophone recognizers, based on hidden Markov models, were tested on

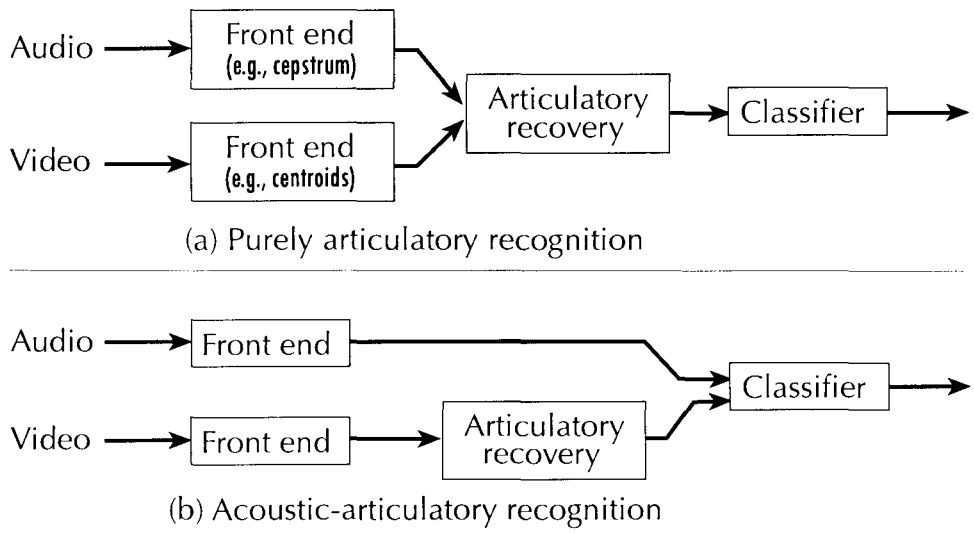


Figure 3.1: Approaches to articulatory recognition, not implemented for this chapter, but illustrated for discussion purposes only. The approach of chapter 2, depicted in detail in figure 2.14, is similar to (b). □

paragraphs each lasting about 20 s. Results were evaluated at the phone level and tabulated by several classes (vowel, stop, and fricative). Measured articulator coordinates were transformed by principal components analysis (projection into the space of the first 4 principal components), and velocity and acceleration were appended.

3.1 Long-Term Goal

Consider the hypothetical recognizer architecture in figure 3.1(a). The computer would accept audio and video inputs of a person speaking. The first stages of processing would be conventional speech-recognition and lipreading front ends. An intermediate stage would recover the motion of the speaker's articulators. *Articulators* include all the parts of the speaker's anatomy that are directly relevant to the acoustics of speech: for example, the tongue, teeth, lips, and vocal cords.

The work described in the present chapter implements figure 3.2 and addresses how well a conventional back end might perform if the articulatory inference functioned precisely. Results obtained here can be viewed as an upper bound for how this particular architecture would perform with motions recovered from sound.

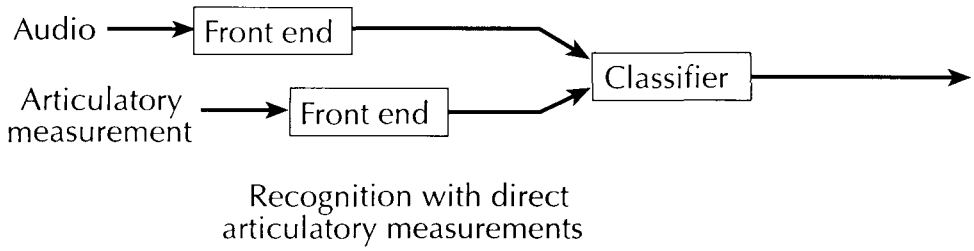


Figure 3.2: The recognition architecture with direct articulatory measurements that was used for the results of this chapter and is shown in more detail in figure 3.6. □

3.2 Problem Statement

The problem addressed in this chapter is, How much can a monophone HMM-based recognizer be improved by adding a representation of measured articulator motions to sound input? The recognition task is continuous speech (paragraphs read aloud), and improvement is measured by the relative reduction in error rate.

3.3 Related Work

3.3.1 Human Speech Perception

The gesture/motor/articulatory theories of human speech perception are described in section 1.2.

3.3.2 Interpolating Acoustic Models of Coarticulation

A primary reason for using motor representations for recognition is to model coarticulation. Nevertheless, it is possible that coarticulation might be modeled acoustically, without explicit representation of articulator motion section 1.1.2. Previous work involved a variant of hidden Markov models that interpolated between acoustic targets (Deng et al. 1992). Each phone had a target; interpolating these targets allowed coarticulation to be modeled with far fewer parameters than the usual approach of triphone modeling.

The above project was inspired by studies of the interaction between consonants and vowels that follow them, in particular the so-called locus property (Delattre et al. 1955) (Liberman et al. 1967). Acoustic trajectory interpolation should not be confused with deleted interpolation, which is used in grammar modeling (Jelinek 1998b).

Another approach to modeling context acoustically is to make the emission PDF dependent on which state was previously occupied (Sitaram and Sreenivas 1997); this increases the number of model parameters required.

3.3.3 Recognition with a Categorical Motor Representation: The Articulatory Feature Model

Other researchers have built an articulatory recognizer using acoustic data without video or kinematic measurements (Erler and Freeman 1996) (Deng and Sun 1994). Their goal is to identify discrete actions that together define a sound. For example, the sound /i/ ("ee") is made by widening the lips and bringing the tongue up and forward. The relative timing of the lip and tongue motions will vary both randomly from trial to trial and systematically due to context.

Their articulatory feature model (AFM) (Erler and Freeman 1996) used seven discretized parameters to represent the articulators' positions (table 3.1). This simplified representation, chosen for ease of implementation, does not describe certain tongue shapes or lip positions. Grooved fricatives such as /s/, lateral sounds such as /l/, and retroflex sounds involve more complex tongue shapes; and labial-dental fricatives (/f/, /v/) involve an excluded lip position.

The AFM was not used for direct recognition—guessing a text string from sound—but was used to score text-string hypotheses. Given a hypothesis (text string), an appropriately connected HMM was created, as during training of a conventional recognizer (Rabiner and Juang 1993). An emission PDF (chap. 1) was trained for each valid combination of AFM state values; static constraints (e.g., tongue tip cannot be behind tongue body) ruled out many state-value combinations. The connectivity of the HMMs was also subject to various dynamic rules; for example, articulators were required to move monotonically between each phoneme and the next.

The AFM HMM experiments used examples of a single subject speaking single words at a time (2694 training cases and 458 test cases). The main finding was that the static and dynamic articulatory constraints improved performance; even so, results were mixed when compared with a conventional baseline recognizer.

Feature	Values					
Voicing	No	Yes				
Velic aperture	Closed	Open				
Lip rounding	Spread	Open	Medium	Narrow	Closed	
Tongue tip Constriction location	Labial	Dental	Alveolar	Palatal		
Tongue tip Constriction degree	Low	Partial	Closed			
Tongue body Constriction location	Alveolar	Palatal	Front-palatal	Back-palatal	Velar	Uvular
Tongue body Constriction degree	Open	Low	Partial	High	Closed	

Table 3.1: Discretized states of articulatory feature model (AFM) (Erler and Freeman 1996). The AFM was developed with acoustic data only, unlike this thesis project which uses direct articulatory measurement. □

3.3.4 Other Recognizers with Categorical Motor Spaces

Finally, a U.S. patent (Sakamoto and Yamaguchi 1992) describes a recognizer that codes vowels according to their place of articulation; this is not articulatory in the sense of the other projects described here, since it involves a much coarser representation.

3.3.5 Recovery of Positions from Sound

A number of projects have attempted to recover articulator configurations from sound without modeling the dynamics of articulator motion; when dynamics are neglected, geometric constraints can still improve performance (Yehia and Itakura 1996).

In one motion recovery project, a lookup table was used to estimate articulator positions from recorded acoustics (Hogden et al. 1993). Articulator motions were measured using the electromagnetic midsagittal articulometer (EMMA) (Perkell et al. 1992), a set of coils attached to the tongue and lips. Sound recordings were converted to sequences of cepstral components. Vector quantization (VQ) was used to group cepstral vectors, and within each VQ category, articulator positions were averaged across all examples. These average positions became the output of the lookup table,

with the acoustic input partitioned according to VQ. Because of the abrupt VQ category boundaries, and single estimated articulator position per category, the recovered motions were not continuous. Recovery was more successful for the tongue than for the lips.

A number of projects have investigated the use of artificial neural networks for recovery of articulator position from acoustics; a review of early efforts appears in a book (Rahim 1994). One such project used the same type of data (x-ray microbeam) as the present project, and investigated recovery for consonants (Papcun et al. 1992). That project found that articulator positions were more predictable where acoustically relevant to the sound being produced. For example, when making a constriction with the tongue near the roof of the mouth, areas of the tongue farther from the constriction were more variable. The critical articulator positions were fairly well predicted, as measured with root-mean-square displacement, by the neural network.

3.3.6 Articulatory Recovery with Kinematic Models

Rather than analyze a short interval of time in isolation, other researchers have made dynamic models for articulatory recovery. Such projects have typically used self-organizing statistical techniques rather than physical principles; examples are neural networks (Bengio and de Mori 1988) (Shirai 1993), genetic algorithms, and dynamically constrained hidden Markov models (Roweis 2000).

The constrained HMM approach uses models with states connected in a lattice. Because states are connected locally within this lattice, the rate of motion from state to state is constrained. Each state represents an articulator configuration, and the emission distribution, as in a conventional recognizer, models the acoustics for the state. Because motion constraints are implemented with model topology, standard training and Viterbi decoding techniques can be used.

Audiovisual databases have also been used to create two-dimensional HMMs having states that are a cross product of position variables (Welsh et al. 1990). The primary difference between this approach and constrained HMMs is that the latter self-organize without position information in the training data.

3.3.7 Incorporating Kinetics into Articulatory Recovery

The accuracy of articulatory recovery can be further improved by extending the kinematic models described above to include a description of forces (kinetics). An enhanced codebook-based scheme (Sorokin and Trushkin 1996) recovered vocal tract shape from sound with explicit physical modeling

(articulatory synthesis), piecewise linear interpolation, and a final optimization step. The combined codebook-optimization approach to inversion had been previously proposed (Atal et al. 1978) and implemented without interpolation (Larar et al. 1988) (Schroeter and Sondhi 1994). The physical model (Sorokin 1992) used generalized coordinates and a system of second-order linear differential equations. Motor control was modeled by fixed damping coefficients and step-function generalized forces. A set of consonant-vowel, vowel-vowel, and vowel-consonant pairs were synthesized with this model, and the resulting articulatory trajectories were partitioned into approximately linear regions. This set of regions formed a codebook for inversion from sound to articulation. The inversion scheme was tested on combined microbeam and sound data for vowels; the articulator positions determined with the x-ray microbeam were transformed into the generalized coordinates of the physical model, and formant frequencies were extracted from the sound. After picking the closest codebook examples and applying linear interpolation, optimization improved the match between measured formants and the physical model's prediction for the hypothetical articulator positions. A related project extended the physical model to fricatives (Sorokin 1994).

3.3.8 Automatic Labeling of Articulatory Events

A previous project (Parlangeau et al. 1996) involved automatically identifying the time and type of various articulatory events, using combined articulatory and acoustic data. That project was intended to facilitate phonetic research, not just automatic recognition. Synchronized recordings of laryngography, nasal airflow, oral airflow, electropalatography, and acoustics were used as input. Event times were defined as times when autoregressive coefficients changed rapidly (faster than a threshold). Events were then classified by a set of hand-coded rules based on the five measurement channels. The overall automatic labeling system agreed with hand labeling 80% to 90% of the time. A speech recognizer that explicitly modeled independent articulatory events (section 3.3.3) might be able to make use of such automatic label information.

3.3.9 Previous Articulatory Recognition with Measured Data

The closest related project to the present was performed simultaneously and independently and is described in section 3.14.

Another similar project used a rescoring network (Blackburn 1997) to choose among recognition hypotheses in an n-best list. That project used the Wisconsin x-ray microbeam data to train a

speech production model that predicted acoustics for each phoneme. After a conventional recognizer produced the n-best list, a sequence of spectra were predicted for each entry in the list. These predictions were compared to the actual acoustics; the discrepancy between predicted and actual acoustics was combined with the recognizer's log-likelihood output score, and the combination was used to reorder hypotheses. Error rates were reduced by 12% to 20% relative to the original recognizer. Search errors are normally present in an n-best list, because Viterbi scores (section 1.11.2) are not identical to forward-algorithm scores (section 1.11.2). Running the forward algorithm on entries in the n-best list can improve recognition, so a list reordered with the forward algorithm might have been a better baseline for comparing performance.

Previous work included building a word-spotting recognizer using just the articulatory part of the Wisconsin data (Roweis 1999). Those experiments did not ask whether there was any advantage to using articulation instead of acoustics; instead, they were intended to confirm that there was sufficient information in the articulatory channels.

Other projects performed speaker-independent recognition from articulation on isolated words (Zlokarnik et al. 1995); and articulatory recognition of vowels using a neural network (Zacks and Thomas 1994).

3.4 Direct Kinematic Articulatory Measurements

3.4.1 Wisconsin X-Ray Microbeam

The articulatory-recognition experiments described in this thesis make use of x-ray microbeam data from the University of Wisconsin (Westbury et al. 1994). To generate this data set, experimenters attached metal beads to several points in subjects' mouths. During speech, bead positions were tracked and acoustics were recorded. A narrow x-ray beam was with feedback control to track the beads' positions. The x-ray microbeam technology is described in greater detail in appendix E.

Using this apparatus for everyday speech-recognition applications is inconceivable. Aside from the problem of x-ray exposure, the beads were attached to the subjects using dental cement. Researchers who acted as subjects themselves reported significant discomfort (Sorokin and Trushkin 1996).

Examples of articulator motion appear in figure 3.4 and figure 3.5. In both cases, the subject started in a resting position and then said a syllable repeatedly. Also plotted are the approximate locations of the roof of the mouth and the back of the throat. The trajectory is plotted for each bead

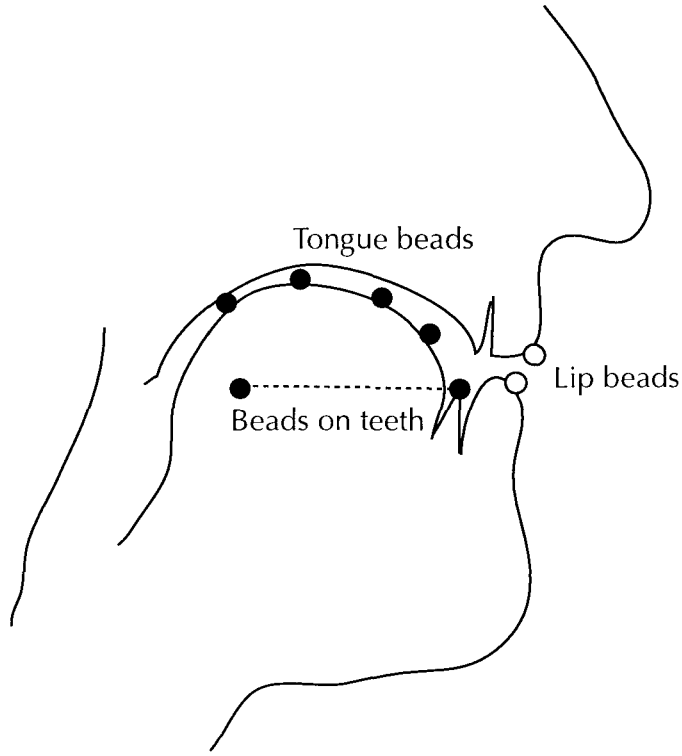


Figure 3.3: Arrangement of beads on head for microbeam recording. Not pictured are three additional beads used to compensate for head motion. This figure also appears as figure A.1. □

affixed to an articulator. In figure 3.4, the subject said "kuh," and as expected, tongue movement is especially prominent. The lower lip moves in figure 3.5 to make the consonant /p/ in "puh."

3.4.2 Alternate Direct Measurement Techniques

Other techniques for directly measuring articulator positions include magnetic resonance imaging (MRI), electropalatography, and cineradiography.

Electropalatography uses an array of electrodes to determine where and when the tongue touches the roof of the mouth or the teeth. A custom acrylic palate, in which the electrodes are embedded, is created to fit each subject. Binary values (contact versus no contact) are recorded for each electrode at each time step. A typical example of array size is 96, and typical sampling rates are 100-200 Hz (Stone 1990) (Wrench 2000).

One study used ultrasound for recovery of the three-dimensional shape of the tongue during steady-state sounds, and compared the results to simultaneous electropalatograhy (Stone 1990). A

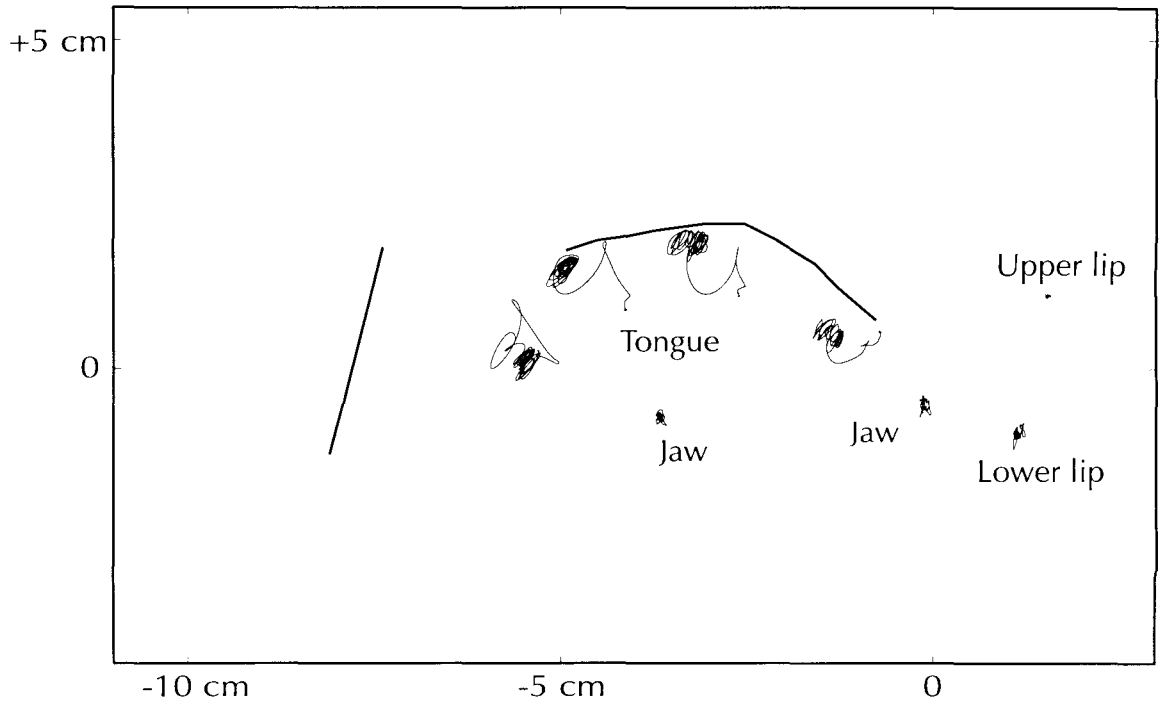


Figure 3.4: Articulator motion as subject says "kuh" repeatedly. Trajectories of all landmark beads (each lip, two on jaw, four on tongue) appear along with the roof of the mouth and back of the throat. For each bead, the dense areas correspond to speech, and the stray traces represent motion from a resting position. □

single subject made a variety of vowels and consonants (excluding stops), sustaining each sound for about 10 s. The researcher's interpretation of the results was that tongue shapes were relatively similar within certain groups of sounds whose regions of contact, determined by electropalatography, were quite different.

Cineradiography involves filming the head with a rapid sequence of x-ray exposures. Due to health concerns, it has not been used for several decades. Fifty-five minutes of old cineradiograph films have been transferred to video and distributed on laserdisc to researchers (Munhall et al. 1994). Despite poor image quality and superposition of different articulators in the image plane, researchers in Japan (Tiede and Vatikiotis-Bateson 1994) and Germany (Höwing et al. 1996) have had some success recovering motions from such data. The Wisconsin microbeam data were better suited to the present thesis work, because they are publicly available, more extensive (longer-duration) than the cineradiograph laserdisc, and include more-direct measurement of landmarks on articulators.

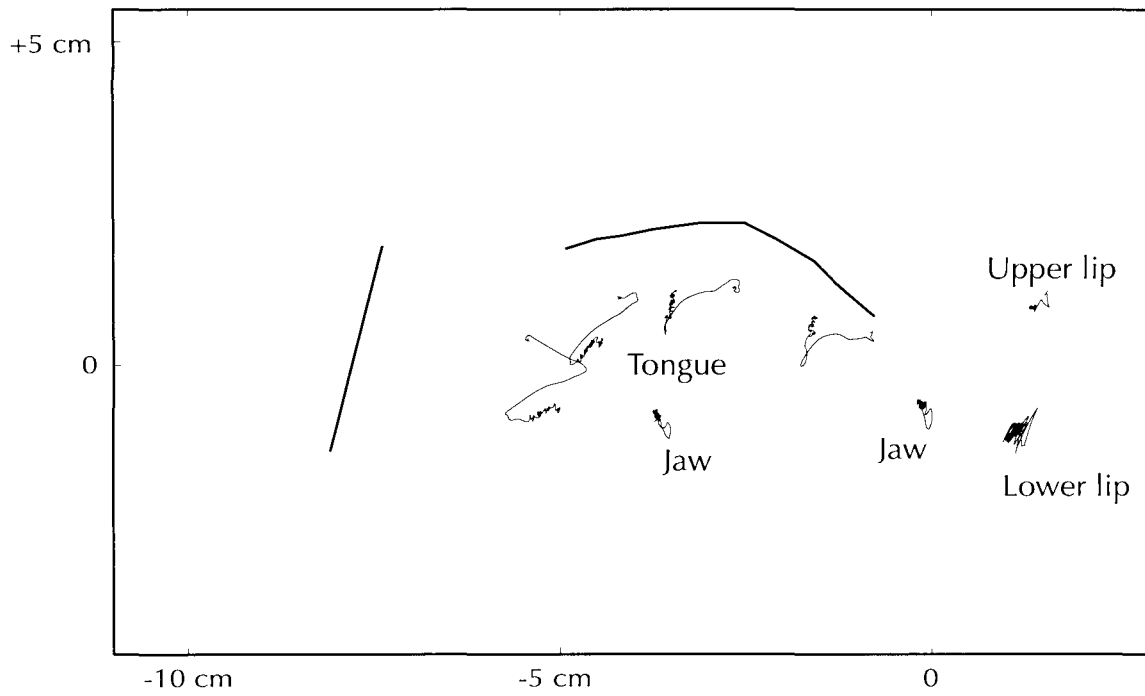


Figure 3.5: Articulator motion for "puh;" see caption for figure 3.4. □

Far greater spatial detail may be obtained with magnetic resonance imaging (MRI) at the expense of time resolution. With MRI, the shape of the cavity formed by the articulators can be directly measured, and predictions of the resonances due to different cavity shapes (Fant 1970) can be tested. Almost all MRI studies have been concerned with steady-state sounds. Early work looked at vowels (Greenwood et al. 1992) (Baer et al. 1991), and a later study measured fricative production (Narayanan et al. 1995). Fricatives were chosen for the latter study because, although they are steady-state sounds, they are not as well understood acoustically as vowels. In recent years, collection times have dropped from order 4 s (Lakshminarayanan et al. 1991) to 0.25 s (Demolin et al. 1997). The faster MRI uses a technique known as turbo spin echo.

A new database for articulatory recognition, Mocha, is being collected at the University of Edinburgh and Queen Margaret University College (Wrench 2000). Unlike the Wisconsin data, which are better suited to speech science than speech technology, Mocha is explicitly designed to be useful for recognizer training. It includes electropalatograph data—a digitized representation of the region of contact between the tongue and the roof of the mouth, sampled on a roughly 8×8 spatial grid every 5 ms. Instead of x-ray microbeam tracking, Mocha uses a set of coils attached to various articulators, which are tracked electromagnetically: an electromagnetic midsagittal articulometer (EMMA)

(Perkell et al. 1992). Nine points are tracked: the upper and lower incisors, upper and lower lips, three points along the tongue, one on the velum, and a reference point on the bridge of the nose. Had Mocha been available in time for the work of this chapter, it might well have been preferable to the Wisconsin microbeam data. In the meantime, data for two subjects have been publicly released.

3.5 Preparation of Data for Recognition

Because of problems in the raw positional data collected at Wisconsin, several preprocessing steps were performed before the recognition experiments. Previous work compensated for subjects' head motions (Westbury 1991) and filled in data lost by errors in bead tracking (Roweis 1999). In the present project, extraneous sounds such as cue tones and experimenters' feedback were removed, and the subjects' exact words were transcribed by hand.

3.5.1 Tracking Correction (Previous Work)

The experimenters in Wisconsin who collected the x-ray microbeam data anticipated that subjects' heads would move during speech. They affixed three reference beads to each subject's head in an attempt to compensate for rigid-body (six degree of freedom) motion of the head. The reference beads were intended to enable after-the-fact correction instead of immobilizing the head during the experiment.

In practice, substantial tracking errors and suboptimal bead placement (Westbury 1991) prevented correction of the full six degrees of freedom. Sagittal translation (up-down and front-back motion) and rotation (nodding) were corrected. Sideways shifting, head shaking, and tilting to the side were uncorrected.

Also, the x-ray microbeam apparatus occasionally lost track of one or more beads. The original data contain the coordinates (1 m, 1 m) for any bead while it was lost.

In a previous project, Sam Roweis established a technique for correcting tracking errors by guessing a bead's most likely position based on the other beads' data (Roweis 1999). The present project starts with corrected data, which he provided for one third of the subjects in the database.

3.5.2 Removal of Extraneous Sounds

The original Wisconsin recordings contain extraneous sounds, primarily cue tones and comments by the experimenter, recorded on the same channel as the subject. Fortunately, almost all such sounds occur at the start (cue tone) or end (experimenter comment) of a recording.

In the present work, every sound file ultimately used in recognition was first opened in an audio editing program to verify transcriptions (section 3.5.4) and adjust the recording duration. New start times—after the end of the cue tone and before the experimenter's comments—were determined. For example, for speaker JW45, the start time was moved forward 689–1852 ms, depending on the file (median 899 ms).

Often the sound recording ended while the subject was speaking. In such cases the end time was moved back to a pause in speech in order to give the file a subjectively plausible-sounding ending.

3.5.3 Total Duration of Each Speaker's Usable Data

After preprocessing, about 10 minutes of data for each speaker were left (for example, 685 s of the subject labeled JW27 in the Wisconsin data). The data for speaker JW45 includes 1673 words. As described in section 3.9.3, most of the recordings were used only for training.

There is no precise theory for the number of examples required to train a speech recognizer. Rules of thumb, based on experience, appear in textbooks and recognizer-development manuals. A standard textbook suggests 40 examples per digit for a five-state, mixture-of-five-Gaussians continuous-digit recognizer (Rabiner and Juang 1993); the manual for the leading software for recognizer development recommends "several hundred utterances" for speaker-dependent monophone models (Young et al. 1997).

The apparent discrepancy between the above suggestions comes from differences in number of models and in the method of counting examples. There are roughly four times as many monophones as digits, and a given utterance referred to in the second guideline above may not include all the monophones. The rules of thumb, then, are reasonably consistent.

Monophones grouped by number of training examples per speaker appear in table 3.2. Because the most common ones occur far more often than the least common, they are grouped into powers of two. Based on the forty-example guideline above, problems with training can be expected for the four phones having less than 2^5 examples. The least-frequently occurring phone, "h," was accidentally combined with "hh," although it should have been merged with "g."

Min	Max	Phones
256	511	ax, n, s, t, r, l
128	255	ih, d, iy, q, k, ah, m, dh, ao, b, ae, eh, z, ow
64	127	p, ux, w, ay, f, hh, aa, v, ey, g, axr
32	63	ng, th, dx, er, y, aw, ch
16	31	jh, oy, uh
8	15	
4	7	h

Table 3.2: Per-speaker phone counts (Westbury et al. 1994), grouped into powers of two. Roughly speaking, the four phones each having fewer than 32 training examples have insufficient data for training. Variability of pronunciation, deviations from the intended text, and truncation of recordings caused some discrepancy for individual speakers. Phones are represented with ARPABET symbols (Shoup 1980) (Rabiner and Juang 1993). □

There are 5500–6200 separate instances per speaker of phoneme-level units in the training data. The variation in the number of instances is due to ambiguity about pronunciation and about whether short silences exist between the words. During training, the system chooses between alternate pronunciations where they exist, and optionally appends a short silence at the end of each word.

3.5.4 Hand Transcription

What the subjects actually said occasionally deviated from what they had been asked to say. Also, as described above, subjects' speech was often truncated. Speech was transcribed by hand so that training data would be correctly described and recognition targets would be accurate.

The hand transcription created in this project used conventionally spelled words, rather than phonemes, allophones, or syllables, because words are easier to objectively identify. The aid of a skilled phonetician was not readily available. As described in section 3.9.4, the recognition system automatically generated phone-like units from the word-level transcripts; this conversion happened during training and used the Viterbi algorithm with monophone models (chap. 1) and an ISO 8859-1 (ASCII) encoded dictionary.

Subject ID	Sex	Age (Years)	Dialect Base (All in USA)
JW27	female	20.9	Blair, WI
JW29	female	20.6	Milwaukee, WI
JW502	female	34.0	Madison, WI
JW12	male	21.1	Marinette, WI
JW15	male	22.4	Milwaukee, WI
JW45	male	21.2	Mishawaka, IN

Table 3.3: Demographic information for subjects used in articulatory recognition. University of Wisconsin researchers who collected the data assigned subject IDs and determined geographical origin of subject's pronunciation. □

3.5.5 Criteria for Inclusion in Recognition Experiments

The following were required for a subject in the Wisconsin data set to be included in the present experiments:

1. Correction of tracking errors available (section 3.5.1; two-thirds of subjects excluded);
2. All six paragraphs extant (to be used, three at a time, for testing);
3. Relatively few truncated recordings (in contrast, some speakers had the majority of their utterances cut short and were therefore rejected for recognition).

Of 48 subjects in the complete data set, few met these criteria, and ultimately six were used for the recognition experiments. The six subjects included three male and three female speakers.

Table 3.3 indicates the sex, age, and dialect base for each subject. The willingness of the University of Wisconsin undergraduates to participate and the requirement that the subject have no metal fillings resulted in a young group. Since subjects were recruited locally, most spoke with a dialect of the Wisconsin region.

There are multiple dialects within Wisconsin. In particular, many urban speakers demonstrate the north inland cities shift (Labov 1996), an ongoing, radical reorganization of the short vowels of English. The severity of the shift can be demonstrated by the following word pairs: a speaker would pronounce "bus" in a way that would sound like "boss" to speakers of other dialects; and "block"

like "black." The "dialect base" judgments previously made by the Wisconsin experimenters are indicated in table 3.3. Subjective listening in the present project suggested that at least one of the speakers—JW27, a twenty-year-old female subject from Milwaukee—used shifted short vowels from time to time. No attempt was made to thoroughly analyze the speakers' dialects, nor to definitively test whether a given subject spoke consistently throughout the experiment. The dictionary also lacked any representation of the shift.

3.6 Recognizer Training

Sound and articulatory measurements were processed by separate front ends. Their outputs were combined and passed to a single back end that was trained using the standard expectation-maximization algorithm. The articulatory front end was trained using principal components analysis.

3.7 Front-End Processing of Sound Recording

The sound recordings of the Wisconsin microbeam data set were processed according to a conventional state-of-the-art procedure. The waveform (sampled at 21.7 kHz) was divided into a sequence of overlapping analysis frames, each lasting 25 ms and with a new frame occurring every 6.87 ms, to match the bead-coordinate sampling rate. For each frame, first 12 coefficients of the Mel-frequency cepstrum were generated with spectral preemphasis. Derivative and energy terms (section 1.10.2, section 1.10.2) were appended, and the resulting feature vector was passed on to the HMM recognizer.

3.8 Front-End Processing of Microbeam Articulatory Data

Microbeam articulatory data were sent through a separate front-end process from acoustics (figure 3.6). The outputs of the two front ends were concatenated at each time step and passed along to the back-end recognizer (section 3.9); the performance of this joint acoustic-articulatory recognizer was then compared to acoustic-only.

For this project, front-end processing of articulatory data included the following steps:

- Coordinate transformation via principal components analysis (PCA).
- Concatenation of first and second time derivatives.

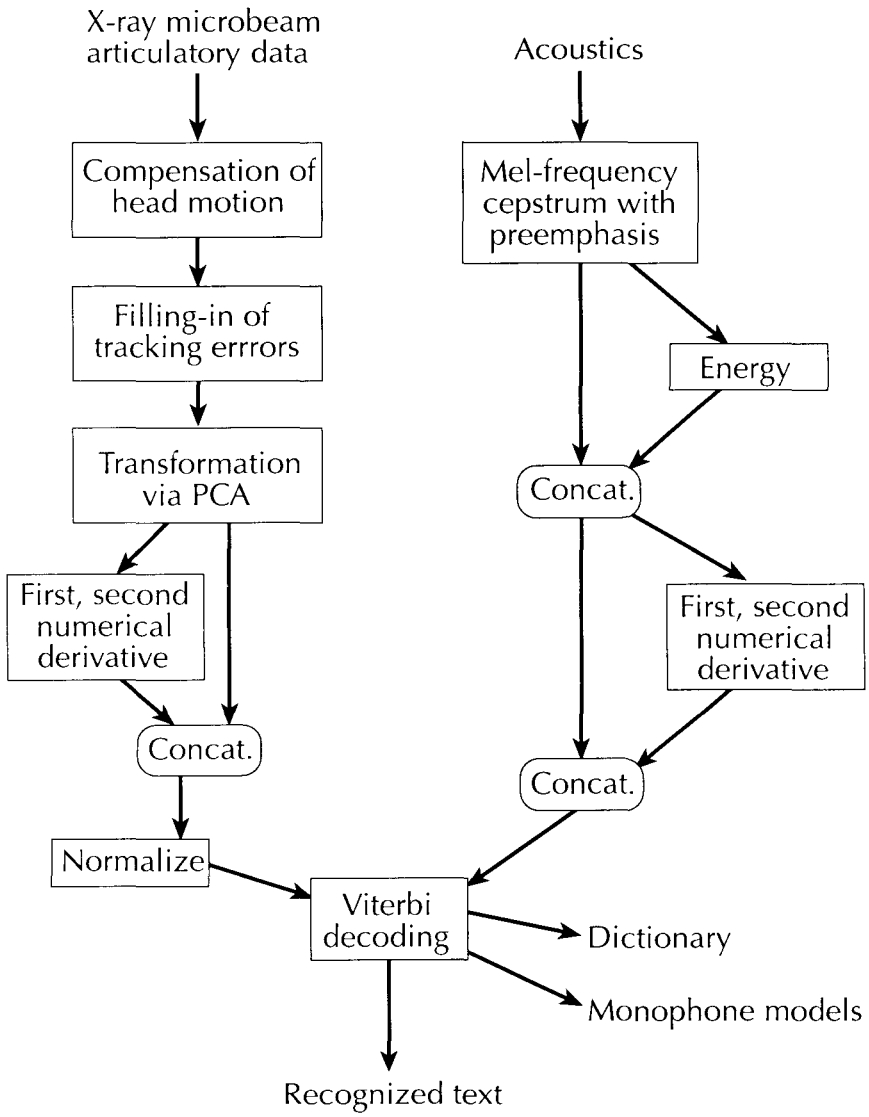


Figure 3.6: Microbeam articulatory recognition architecture, used for experiments described in this chapter. Concatenating microbeam data to acoustics improves recognition (section 3.13). □

- Between-channel normalization of mean and variance over time, where a channel is a front-end output vector component.

The data fed into the above pipeline already had tracking errors corrected as described in section 3.5.1.

3.8.1 Articulatory Parameterization and Constraints

The motions of the articulators are constrained by physiology and phonology. The two beads attached to the jaw are subject to the severe physiological constraint that the jaw is rigid. An example of a phonological constraint is that sticking the tongue out past the lips is not required for making any sound of English. The phonetically-relevant lip shapes illustrated in figure 2.1 represent only a subset of possible configurations.

The constraints on articulator movement limit the number of independent degrees of freedom as well as the range of values they can have. For example, the lower jaw's motion might be described by the angle of its opening as well as the extent of its protrusion, rather than the two coordinate pairs for the two attached beads. Such a transformation would reduce four parameters to two.

3.8.2 Principal Components Analysis (PCA)

For the present project, the raw bead coordinates were transformed into a smaller number of parameters. For most experiments the new parameter set was obtained through principal components analysis (PCA). This technique is defined an analyzed in appendix A.

In figure 3.7, the percentage of total microbeam coordinate variance explained by each successive principal component is plotted for each speaker. For all speakers combined, the cumulative variance explained is presented in table A.1. The first four components explain almost all the variance. This is borne out by a comparison of recognition results for one speaker with 2, 4, 6, or 8 principal components used for recognition (section 3.12.2).

For recognition, the n_{PC} directions having the greatest variance were kept, while the other $(16 - n_{PC})$ were discarded. The bead coordinates $\mathbf{x}(t)$ for each sample time were transformed by taking the dot product of the coordinate vector with each of the principal components having the largest n_{PC} eigenvalues:

$$\alpha_k(t) = \mathbf{p}_k \cdot \mathbf{x}(t)$$

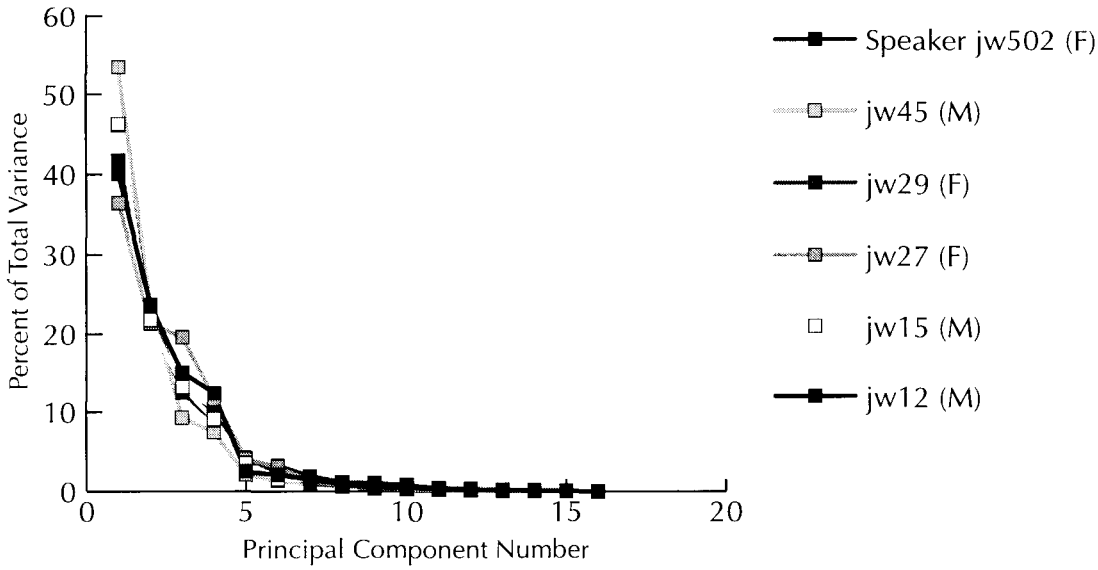


Figure 3.7: Relative variance of principal components, suggesting that most information is retained by the first 4–8 components (when used as coordinate bases). The sex of each speaker is indicated, since recognition results differed by sex. □

The resulting features, $\alpha_k(t)$ for $1 \leq k \leq n_{PC}$, are then differentiated and normalized across different times t.

3.8.3 First and Second Time Derivatives

After microbeam data were reduced to n_{PC} parameters for each sample time, the first and second time derivatives were added for each transformed parameter. These derivatives are analogous to the velocity and acceleration of the individual beads; however, since they are computed in the transformed space, they describe the coordinated motion of multiple beads.

Adding the derivative parameters was partly inspired by the success of acoustic derivative parameters for recognition (Fry and Denes 1958) (Denes and Matthews 1960) (Furui 1986) (Rabiner and Juang 1993). These acoustic time derivatives, often called delta and acceleration coefficients, are computed from the sequence of cepstral vectors (chap. 1).

3.8.4 Normalization

The articulatory and acoustic parameters were normalized across all utterances so that different parameters would have the same average and variance over time. The primary motivation for this

normalization was so that the variance floor (section 3.11.1) would be appropriate for all the parameters.

3.9 Back-End Recognizer for Acoustics and, Optionally, Articulation

The back-end recognizer was based on a set of monophone CDHMMs (section 1.11.2) and a unigram model for grammar (section 3.9.1). Training began with a flat start (section 3.9.4) and followed a sequence of E-M iterations, retranscriptions of the training data to better match actual pronunciation, and splitting Gaussians in each state's emission PDF (section 3.9.2). Unfortunately, there is no theoretical way to determine the optimal sequence of operations, so the sequence was established through trial and error. In addition to the many model parameters that were updated via EM (section 1.11.2), three global parameters affecting all models were optimized by repeating experiments (section 3.11).

Experiments were always run in pairs, representing two different splits of data into train and test (section 3.9.3). For each experiment in a pair, training examples never appeared in the test set—except for the results reported in section 3.12.4, which compares performance on the two types of data.

The number of model parameters for recognition was determined by gradually increasing the complexity of the models and noting the point at which generalization ability deteriorated (section 3.12.2).

A total of 45 models were trained for each speaker, for the 43 monophones and two types of nonspeech intervals. The latter models are often, in other projects, referred to as silence models.

3.9.1 Unigram Model for Grammar

In this project, prior probabilities for words were approximated with a unigram model, in lieu of a grammar model. The probability estimates $\hat{P}(w)$ were obtained by dividing the number f_w of occurrences of each word w by the length L, in words, of the transcript:

$$\hat{P}(w) = \frac{f_w}{L} : f_w \geq 10$$

Words having fewer than 10 occurrences were all given the same probability estimate, which was equal to the total number of instances of all subthreshold words, divided by the number of distinct subthreshold words.

Count	Words
209	*interword silence*
100	the
55	a
36	of
29	in
23	all
22	had, to
20	one, that
19	two
18	and
17	he
15	blend, five, house, school, sense, special, things
14	his
13	four, is, three
12	back, eight, I, seven, six
11	coat, long, nine, people, problem, you
10	across, cash, children, country, dark, dormitory, light, make, moment, much, nothing, order, point, row, second, ship, shoot, street, told, wax

Table 3.4: Word frequencies used to create unigram model. Only words with ten or more occurrences were included; all words falling below that threshold were given an equal share of their combined frequencies. □

$$\hat{P}(w) = \left(\frac{1}{L}\right)\left(\frac{\sum_i f_i}{\sum_i 1}\right) : f_w < 10. f_i < 10$$

3.9.2 Training Sequence for Articulatory Recognizer

The recognizers for this project were trained by the following procedure. Section 3.9.6 reports recognition results for some variations that were rejected.

- Identify usable speaker; preprocess transcriptions, sound, and microbeam recordings as described in section 3.5.
- Split data into training and testing sets (section 3.9.3).
- Flat start (section 3.9.4):
  - Initialize monophone HMMs with identical parameters.
  - Perform a forced alignment using the untrained HMMs.
- Training iterations:
  - Reestimate model parameters.
  - Optionally, increase model complexity (number of Gaussians in emission distributions).
  - Optionally, test recognition performance (section 3.10).
- Segment recordings into monophones using trained models (section 3.9.5).
- Initialize a new set of monophone models using segmented recordings.
- Repeat training iterations with new models.

3.9.3 Train/Test Split

When data were split into training and testing sets, some recordings were discarded: nonspeech examples (e.g., swallowing) and awkwardly modified speech (e.g., in response to a request to speak extremely slowly). Each recognition experiment was repeated for two different train/test splits.

For a given split, the text of the train and test sets would have been identical across subjects, had the Wisconsin data collection gone perfectly. However, because of problems such as truncated recordings (section 3.5.2) and mistakes by the subjects (section 3.5.4), there were slight differences between subjects.

In the first split, the test set included tasks 11, 78, and 80 of section F.1. The second split used tasks 12, 79, and 81 for testing. In each case, the paragraphs not used for testing were included in the training set.

Paragraphs were chosen for testing because of their duration, and because the other utterances—words and sentences—were repeated throughout the data set in a way that made them difficult to separate (table 3.5). Each sound file (except for paragraphs) contained several utterances, and

- Task 10. "*The other one is too big.*
  Don't do Charlie's dirty dishes.
  She had your dark suit in greasy wash water all year."

- Task 45. "*The other one is too big.*
  She always jokes about too much garlic in his food.
  If I had that much cash I'd buy the house."

- Task 46. "The point of the program will be told before long.
  Across the street stands a country school.
  She had your dark suit in greasy wash water all year."

Table 3.5: Examples of sentence tasks in microbeam data set. The data are difficult to partition into distinct train and test sets, because sentences are repeated across tasks and grouped within each task. Instead, paragraphs (appendix F) were used for testing. The repetition of sentences also makes training difficult by reducing the number of contexts in which each phoneme appears. □

the grouping was permuted so that a given sentence/word would be spoken in several different combinations. If sentences had been used, either the recognizer would have seen several training utterances of each test sentence, those redundant utterances would have been thrown out (along with the other utterances in their sound files), or the data set would have needed to be divided into individual sentences and words.

The test splits described above still left some redundancy between train and test data, because the paragraphs actually covered overlapping regions of the original text (see appendix F). Four of the paragraphs had overlap at one end (either start or end), while two overlapped at each end. If overlap were causing a significant improvement in performance, the doubly overlapping paragraphs should be easier for the recognizer than the singly overlapping examples. Contrary to this prediction, the error rate for doubly overlapping cases (¶4 and ¶5 in table 3.9) was no lower than for singly overlapping ones (¶1–¶3 and ¶6).

3.9.4 Flat Start

The flat-start training procedure (Rabiner and Juang 1993) is used when training data have not been previously segmented into phoneme-level units. In the present project, as is often true, not only are

monophone start and end times lacking, but each recording was transcribed as a sequence of words rather than monophones. In this case, arbitrary pronunciations of each word are selected to create a monophone transcription. The flat start involves initializing all monophone models with the same parameters.

3.9.5 Restarting Training with New Models

Theoretically, information about start and end times of monophones should be useful for initiating training—starting with segmentation should outperform a flat start.

In this project, the models that originated in the flat start were used to create a segmentation. New models were then initialized and training was restarted. The new models started with only one Gaussian in each emission distribution, while the old models had grown to have 16 Gaussians in each distribution by the time they produced the segmentation. Parameters were gradually added to the new models until they reached the same level of complexity as the old ones.

3.9.6 Optimality of Recognizer Architecture

Recognition results verified that restarting training as described above improved performance over a flat start. For example, the average error rate (across all speakers) of acoustic-only recognition was $(29.7 \pm 1.6)\%$ with the models trained from a flat start; when these models were used to segment and initialize a new set, the average error rate dropped to $(27.83 \pm 1.5)\%$ (after retraining).

3.10 Recognizer Testing

Each recognition experiment involved repeated recognizer testing throughout the training procedure. Testing was performed for different untrained parameters (grammar weight and insertion bias) as well as different degrees of model complexity (number of Gaussians in each emission distribution). Testing was also performed before and after the flat-start models were discarded (section 3.9.5).

The motivation for such extensive testing was to optimize both recognizers—acoustic-only and joint acoustic-articulatory—to ensure the fairest possible comparison between the two.

Speaker ID	Sex	Acoustic Only Error Rate	Joint Error Rate	Relative Improvement
JW27	female	29.7%	30.0%	–0.9%
JW29	female	28.4%	26.9%	5.2%
JW502	female	20.6%	20.0%	2.8%
JW12	male	28.3%	25.4%	10.4%
JW15	male	30.5%	26.4%	13.5%
JW45	male	29.5%	25.2%	4.7%

Table 3.6: Comparison of joint acoustic-articulatory to acoustic-only recognition, for each of the six speakers. Error rates are for monophone sequences, and articulatory improvement is reported as a fraction of the acoustic-only error rate. The improvement is statistically significant across all speakers ($p = 0.018$) and for males considered separately ($p = 0.006$) but not for females considered separately ($p = 0.36$). □

3.11 Setting Global Parameters of Recognition

Several global recognizer parameters, described in more detail below, needed to be determined. Grammar weight and insertion bias were jointly optimized during testing. Only a few settings were tested for the third parameter, variance floor. To test it completely would have increased the search space combinatorially, and required not just retesting but also retraining for each new setting; experiments were already using considerable processing power and time.

3.11.1 Variance Floor

The variance floor (Rabiner and Juang 1993) works around the problem of estimating many parameters (in a mixture-of-Gaussians CDHMM) from relatively little data. With large datasets, this problem arises when context-dependent (biphone or triphone) models are used. With small datasets, such as in the present project, parameter estimation is a potential problem even for monophone models.

For each Gaussian in each state's emission PDF, no covariance matrix component is allowed to drop below the variance floor during training. In the case of diagonal covariance matrices, the off-diagonal elements are still set to zero.

The units of the variance floor parameter are poorly defined. Each covariance matrix component

represents a product of two front-end output components. Even with a purely acoustic front end, this causes problems, because cepstral coefficients appear alongside derivatives and energy. Kinematic components introduce completely different units. Restricting the covariance matrix to diagonal elements does not solve the problem, because the same floor is typically applied to all the elements.

In the present work, the problem of ill-defined units is addressed by normalizing each component of the feature vector to have zero mean and unity variance over time.

Relative to each component's overall variance of 1, the variance floor was set at 0.01. Other tested values, which underperformed the chosen setting, were 10^{-4}, 0.005, 0.05, and 0.25.

3.11.2 Grammar Weight and Insertion Bias Defined

Due to the inadequacies of the acoustic model and the grammar model, one of these models will be more reliable in recognition than the other. The grammar weight g adjusts the relative importance of the acoustic probability density p_A and the grammar probability P_G for every output hypothesis λ of length L (equation 3.1) (Young et al. 1997). Although it is theoretically unsatisfying (Bourlard et al. 1996), it is used ubiquitously in the best-performing recognizers (Jelinek 1996).

Insertion bias b, also in equation 3.1, specifically compensates for a possible bias of the grammar model toward shorter-length strings (Young et al. 1997).

$$p(v, \lambda) = e^{bL} \left(\frac{P_G(\lambda)}{L} \right)^{gL} p_A(v|\lambda)$$

Equation 3.1: Definition of grammar weight g and insertion bias I. □

The calculation is carried out in the log domain:

$$\ln p(v, \lambda) = bL + gL \left(\ln P_G(\lambda) - \ln L \right) + \ln p_A(v|\lambda)$$

3.11.3 Joint Optimization of Grammar Weight and Insertion Bias

For each speaker and each train/test split, recognizer testing was repeated for 20 different combinations of parameters; four settings of insertion bias and five of grammar weight were jointly tested. Performance as a function of the two parameters is displayed graphically in figure 3.8, in which the error rate is plotted on an inverted scale.

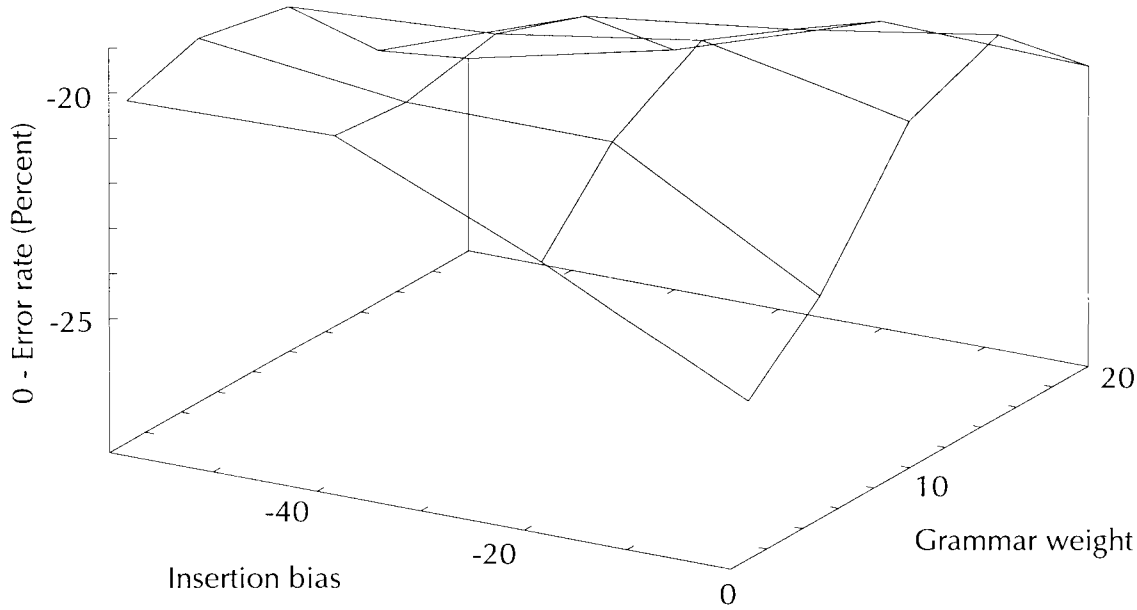


Figure 3.8: Error rate (inverted scale) as a function of grammar weight and insertion bias for speaker JW502 in the Wisconsin data. For this graph, error rates were averaged across the two train/test splits; elsewhere, cross-validation was used (see text). These parameters were optimized for recognition, but a range of settings give comparable results. □

Ideally, these parameters would have been optimized on a development test set and tested once on an evaluation test set. This would have ensured that the recognizer generalized correctly to new data and that after-the-fact parameter setting was not critical to performance. Due to limited data and time, strict separation and single-pass evaluation were not possible. However, a form of cross-validation was used, in which the optimum grammar weight and insertion penalty for a particular train/test split (section 3.9.3) were used to evaluate the other of two splits. An example illustrating this procedure appears in table 3.7.

3.11.4 Comparison to Other Recognizers

The phone error rates obtained in the present project are similar to the performance of one state-of-the-art recognizer on the same data and to other recognizers on the TIMIT data set (Garofolo et al. 1993). However, those recognizers are all speaker-independent, and trained on much larger data sets, so the rates cannot be precisely compared.

The paragraphs for speaker JW45 were presented to the commercially available IBM ViaVoice

	Split 1				Split 2			
Grammar	Insertion bias				Insertion bias			
weight	-60	-40	-20	0	-60	-40	-20	0
1	22.2	21.2	23.3	26.2	18.6	19.4	21.2	22.8
5	21.8	22.2	22.1	24.9	18.0	18.8	18.9	21.2
10	21.1	21.6	**21.2**	21.6	19.6	18.6	*17.5*	19.0
15	22.5	21.6	21.2	*20.1*	22.4	20.0	19.3	**18.9**
20	22.6	21.8	21.5	21.6	24.9	23.5	20.9	21.1

Table 3.7: Example of cross-validation of grammar weight and insertion bias, used to calculate error rates throughout the rest of this chapter and chapter 4. For each train/test split (section 3.9.3), the optimum value is indicated in italics. These parameter settings are then used to determine the error rate, indicated in boldface, for the other train/test split. □

Gold recognizer, which transcribed them with a phone error rate of 26.3% (29.9% on split 1, and 22.7% on split 2) (section 3.9.3). Since speaker enrollment for that recognizer requires the user to read text prompts interactively, it was not possible to adapt the recognizer to the speaker. The vocabulary was not constrained to words in the Wisconsin test set; restricting the vocabulary would have also aided recognition. A breakdown of ViaVoice's errors by phonetic class—vowel, stop, and fricative—appears in section 4.5.

3.12 Increased Recognition Performance with Measured Data

3.12.1 Recognition of Consonants versus Vowels

Recognition results were examined for different sound classes (vowel, stop, and fricative). Chapter 4 describes these results in detail, compares them to conventional recognizer errors, and includes a preliminary investigation of the information conveyed by each sound class. Fricatives had the lowest error rate, followed by vowels, with stop consonants having the highest. Inclusion of microbeam data improved recognition of fricatives and vowels much more than stops.

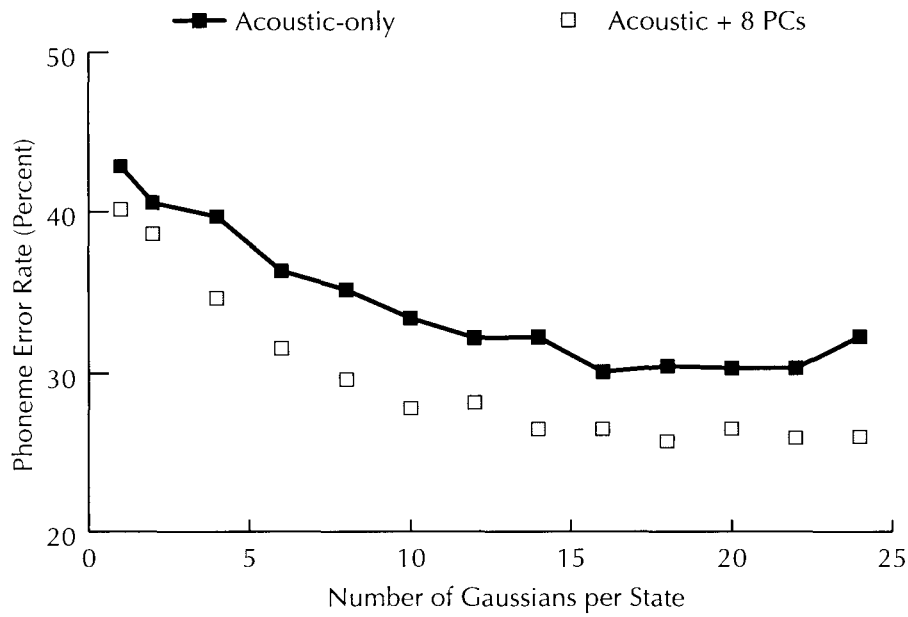


Figure 3.9: Test-set error as training progresses and parameters are added. Most of the results reported in this chapter use 16 Gaussians per state. □

3.12.2 Number of Parameters Required

Experiments addressed the number of model parameters and front-end features required for optimal recognition. Model parameters were increased by adding modes to the emission distributions in the HMM. These modes were additional Gaussians added to the mixture model for each state. Results are shown in figure 3.9; with or without microbeam channels, near-optimal performance is achieved with 16 Gaussians per mixture.

3.12.3 Variability between Test Paragraphs

Of the six paragraphs used for testing (appendix F), some were considerably easier to recognize than others (table 3.9). The error rates varied by roughly a factor of 2.

3.12.4 Performance on Training Data

Throughout this chapter, error rates are reported for test partitions of the data. This is the preferred way to measure performance because it requires the recognizer to generalize from training examples to unseen data. For completeness, table 3.10 reports recognizer testing on the paragraphs in the

	Flat Start	Realigned
Acoustic-only	32.4%	29.5%
4 PCs	27.9	28.1
4 PCs + Δ + Δ^2	27.9	26.5
6 PCs + Δ + Δ^2	27.9	N.A.
8 PCs + Δ + Δ^2	28.1	N.A.

Table 3.8: Phoneme error rate as a function of articulatory parameters used: number of principal components, and whether velocity and acceleration were included. Cases marked "N.A." were not tested. □

	¶1 Error Rate	¶2 Error Rate	¶3 Error Rate	¶4 Error Rate	¶5 Error Rate	¶6 Error Rate
Acoustic-only	27%	33%	34%	39%	37%	23%
4 PCs	26	24	29	32	35	21
4 PCs + Δ + Δ^2	21	22	28	40	33	20

Table 3.9: Phoneme error rate on individual paragraphs in the test set, with or without velocity and acceleration, and with or without articulation □

Acoustic-only	8.91%
8 PCs + Δ + Δ^2	8.68

Table 3.10: Phoneme error rate on paragraphs in training set, for speaker JW45 □

training set. This experiment was performed for only a single speaker—JW45. As expected, the recognizer performs far better on the training paragraphs; the error rates for that set are about one third of the test-set rates.

Articulatory information helps recognition of the training paragraphs, but only slightly. Perhaps the articulatory representation lends itself to generalizing between seen and unseen data. Testing this hypothesis properly is outside the scope of this thesis.

The recognizer used to generate table 3.10 is essentially like the one reported in table 3.6 for

speaker JW45; however, the preprocessing steps of differentiation and normalization were swapped because of an implementation error. In the former recognizer, articulatory parameters (transformed by PCA) were normalized before first and second time derivatives were computed. In the latter, the first and second derivatives were concatenated to PCA output before normalization occurred. The normalization set the time average and variance equal across channels, as described in section 3.8.4.

3.13 Summary of Results

The experiments described in this chapter demonstrated that adding direct measurement of articulator motion to acoustics, in a conventional HMM architecture, resulted in a relative reduction of 7.4% in the phone error rate. Articulatory information had a greater impact on vowel and fricative recognition than on stop consonants, by about a factor of 3 in relative error rate. The improvement due to articulation was statistically significant (and greater) for the three male speakers studied but not for the three female speakers.

3.14 Recognition with Microbeam Data: Conclusions and Discussion

The present project demonstrates, with across-speaker statistical significance for the first time, an improvement in continuous-speech recognition when direct articulatory measurements are concatenated to acoustics.

A simultaneous, independent project built an analogous recognizer for electromagnetic midsagittal articulometer (EMMA) data (Wrench and Richmond 2000). That project tested statistical significance separately for only two subjects, unlike the present work which used a cross-speaker analysis to test generalization. Additional analyses are also presented here, showing differing results for male and female subjects and tabulating results for different phonetic classes.

Previous work has formulated articulatory models without actual data (Erler and Freeman 1996), added articulatory modeling as a postprocessing step to conventional recognition (Blackburn 1997), used isolated fragments of speech (Papcun et al. 1992), or performed articulation-only recognition without comparing it to acoustic (Roweis 1999).

It is reasonable that microbeam articulatory data improved recognition more for vowels than for stop consonants. Previous work, which looked at phonemes in the context of words, found that a

difference in the manner of production of consonants distinguished two words more often than a change in the place of production (Denes 1963). This observation is applicable to the present work because the latter uses a dictionary, rather than a phone-loop recognizer. It means that consonant place is more redundant than consonant manner, in that place is easier to predict from the rest of the word. The positional data—neglecting delta and acceleration coefficients—directly represent place of production. A similar phenomenon has been observed for audio and video, in which audio has a greater influence on perceived manner of articulation, and video on perceived place (MacDonald and McGurk 1978).

3.15 Future Work

The present results suggest the following next steps:

1. Test the recognizer using articulatory positions reconstructed from acoustics, rather than directly measured. Such positions might be obtained using constrained HMMs (section 3.3.6), which have been demonstrated to work for the shorter utterances in the Wisconsin data set (Roweis 1999).

2. When an adequate number of speakers become available, work with the Mocha data set (section 3.4.2). Some preliminary recognition results were recently reported by the group that collected the data (section 3.14) (Wrench and Richmond 2000).

3. Apply more-sophisticated parameter extraction, such as the heteroscedastic technique discussed in appendix B.

4. Instead of introducing articulatory data to a conventional recognizer architecture, as described in this chapter, use the data to train recognizers that attempt to model articulation (section 3.3) but which have only been trained using acoustic data.

3.15.1 Testing with Inferred Articulation

The recognizer developed in this chapter could easily have used kinematic information inferred from acoustics. The constrained hidden Markov model technique has been demonstrated to work on short utterances in the Wisconsin data (Roweis 1999). Assuming it scales up to longer files—the paragraphs—it should show similar improvement in recognition rate over acoustics alone. This would

be a direct demonstration of the effectiveness of an intermediate motor representation for speech recognition. The separate demonstrations of the two stages of articulatory recognition—motion extraction and recognition from motion and sound—are nevertheless valuable.

3.15.2 Applying Optimal Feature Extraction

A straightforward and promising next step is to apply heteroscedastic feature extraction (HDA, appendix B) instead of principal components analysis (section 3.8.2; section B.4.1) to the kinematic representation. The case for using HDA on kinematic parameters is stronger than for using it with the acoustics, since acoustic front ends and conventional back ends have been jointly improved by many researchers over the years.

Chapter 4

Recognition and Entropy of Vowels and Consonants

It is often assumed that mainstream speech-recognition systems perform better on vowels than consonants. For example, a leading textbook claims,

> The vowel sounds are perhaps the most interesting class of sounds in English. Their importance to the classification and representation of written text is very low; however, most practical speech-recognition systems rely heavily on vowel recognition to achieve high performance.... (Rabiner and Juang 1993)

The book does not quantify these claims, and it is rare, in the literature, for recognition results to be tabulated by broad phoneme classes. The present study considered separate classes of consonants; it empirically tested whether the vowels convey less information in text and whether several speech-recognition architectures actually perform better on them.

When only stops and fricatives are contrasted with vowels, and a phonetic transcript is used, the claim about vowels' importance no longer applies. The discrepancy, by the implied above quote, between vowel and consonant recognition is contradicted here by error analysis of several recognizer architectures.

4.1 Different Types of Consonants

The stop, fricative, and affricate consonants are particularly interesting because they are the least similar to vowels. Other consonants, such as /l/, /m/, /n/, /r/, and most obviously /y/ and /w/, are closely akin to vowels. Following is an example of a sentence with letters deleted except for stops, fricatives, and affricates:

Th__ __t_d s_g__f_c__t __p__v____ts __ th_ c__p___'s ___g_,...

The textbook quoted above uses the same example, but includes all consonants:

Th_y n_t_d s_gn_f_c_nt _mpr_v_m_nts i_ [sic] th_ c_mp_ny's _m_g_,... (Rabiner and Juang 1993)

Here are two more example sentences, one with stops and fricatives retained, and the second with only vowels retained:

G___d __s c___ f__ ___ __d s___ __ __s s____.

_e __a_ _e__ i_ __o__ a__ _i__ a__ __o___.

The sentences appear in full in appendix F. As these examples suggest, excluding vowel-like sounds from the consonant category evens out the information conveyed by vowels and consonants. When phonetic transcription is used instead of standard English orthography, the two categories have almost identical information content (figure 4.1).

The scope of the consonant category matters less to the recognition results, because the recognition rates are averages across all the phones in the category, while entropy is an ensemble property that increases with ensemble size.

4.2 Entropy of Vowels and Consonants

Some rough estimates of the information content of vowels, stops, and fricatives were generated. The estimates were derived from English text, English phonetic transcription, and Basque text. Basque was included to test whether the findings were language-specific, because it is a non-Indo-European language, almost totally unrelated to English. Basque was chosen also because it uses Roman letterforms, making it easy to process. Symbol frequencies were used to compute the entropy of vowels, stops, and fricatives.

Symbol frequencies were obtained from the following sources:

1. English text: a Web page listing letter and bigram frequencies in Charles Dickens' *A Tale of Two Cities* (Hahn 1994).

2. English phonetic transcription: phone frequencies listed in the documentation accompanying the Wisconsin microbeam data set (Westbury et al. 1994).

3. Basque: analysis of a 7312-character Web page with tourist information on the Basque region. This small sample cannot be used for precise comparison, but suggests that results are not radically different for this non-Indo-European language.

4.2.1 Categorization of Vowels, Stops, and Fricatives

Finding occurrences of vowels, stops, and fricatives was straightforward for the phonetic English and for Basque; the conventional English text presented problems.

Standard English Spelling, Dickens Sample

Because English spelling is not phonetic, the letters of English were heuristically classified for this project. Fortunately, the source for letter frequencies in Dickens (Hahn 1994) also included bigram (letter pair) statistics. Bigram frequencies were used to differentiate between hard and soft "c:" in "ce" and "ci," the first phoneme was assumed to be the sibilant /s/, and in other contexts, "c" was equated with the stop /k/. Although this rule is not universally applicable, it prevents egregious misclassification of "c." Similar bigram rules were used for "g," "th," and "ng" (table 4.1).

The letter pair "th" presented problems: it may represent /ð/ (as in "that"), /θ/ (as in "south"), or the sequence /t h/ (as in "anthill"). The former two cases were lumped together, and the third case was neglected.

No simple rule could precisely categorize "ng." The letter pair can signify /ŋ/ ("thing"), /ŋg/ ("anger"), or, particularly across word boundaries, /n g/ ("on guard"). All occurrences of "ng" were treated as /ŋ/, which neglected /g/ and reduced the frequency estimate of stop consonants.

English Phonetic Transcription

Classifying symbols in the Wisconsin transcription was straightforward, because it was phonetic (table 4.2). This transcription was idealized—it didn't represent the errors subjects and experimenters made in individual trials—and was not used in the recognition experiments of chapter 3.

Vowel	a, e, i, o, u
Stop	b, c, d, g, k, q, *t*
Fricative	c*e*, c*i*, f, h, s, th, v, z
Other	ch, g*e*, g*i*, h, l, m, n, ng, r, y

Table 4.1: An approximate grouping of English letters and letter sequences into vowel, stop, and fricative categories, used to generate the results of figure 4.1. Specific letter combinations overrode the default classification of one (e.g., the "c" in "ce") or both (e.g., "n" and "g" in "ng") symbols. □

Vowel	aa, ae, ah, ao, aw, ax, axr eh, er, ey, ih, iy, oh, ow, oy, uh, ux
Stop	b, d, dx, g, t, k, p, q
Fricative	f, s, sh, th, z
Other	h, hh, l, m, n, r, w, y

Table 4.2: Phonetic symbols grouped into vowels, stops, and fricatives. These symbols were used in the phonetic transcription and phone frequencies provided in the Wisconsin x-ray microbeam data. □

Vowel	a, e, i, o, u
Stop	b, d, g, k, p, t
Fricative	f, h, s, x, z
Other	*c*, j, l, m, n, r, *v*, *y*

Table 4.3: Categorization of Basque letters into vowel, stop, fricative, and other. The letters "c," "v", and "y" are not native to Basque, but are used in writing foreign words. In the sample text of 7312 characters, these number of occurrences of these letters were seven, one, and one, respectively. □

Basque Text

The Basque letters were also easy to classify (table 4.3), since spelling in that language is much more phonemic than English. Although it uses Roman letterforms and phonemic correspondences, it is essentially unrelated to other Indo-European languages. The categories were derived from a guide to Basque pronunciation for English speakers.

4.2.2 Measurement of Entropy

Entropy, in the information-theoretic sense, can be used to measure the unpredictability of letters in a natural language (Shannon 1951). For example, if an English text is split at an arbitrary point, and only the first half is presented to an observer, what can be inferred about the first letter of the second half? There are many ways of estimating the entropy of English, including having subjects place actual bets on what letter will appear next. This gambling approach yielded an estimate of 1.3 bits per symbol (Cover and King 1978). More recently, an upper bound of 1.7 bits per symbol has been established (Brown et al. 1992). Generally these studies do not consider vowels and consonants separately, except for an early study that counted the frequency of commonly-occurring symbol combinations, but did not calculate entropy (Denes 1963).

4.2.3 First-Order Entropy Estimation

A crude upper bound on entropy, used in this project and given in the following equation, is first-order uncertainty based on symbol frequencies. It is an upper bound because the preceding context of symbols in English helps to predict what symbol will come next.

$$I = -\frac{1}{N}\sum_{\sigma} F(\sigma)\log_2\frac{F(\sigma)}{N}$$

Equation 4.1: Entropic information I from symbol frequencies $F(\sigma)$, measured in bits □

The above quantity can be written as a sum of the contributions of different subsets of symbols:

$$I = I_V + I_S + I_F + I_O$$

The subscripts V, S, F, and O represent the analysis categories of vowel, stop, fricative, and other. Each term has the same form; the following is the contribution of vowels:

$$I_V = -\frac{1}{N}\sum_{\sigma\in V} F(\sigma)\log_2\frac{F(\sigma)}{N}$$

The percent of total entropy due to the vowels, plotted in figure 4.1, is $100(I_V/I)$.

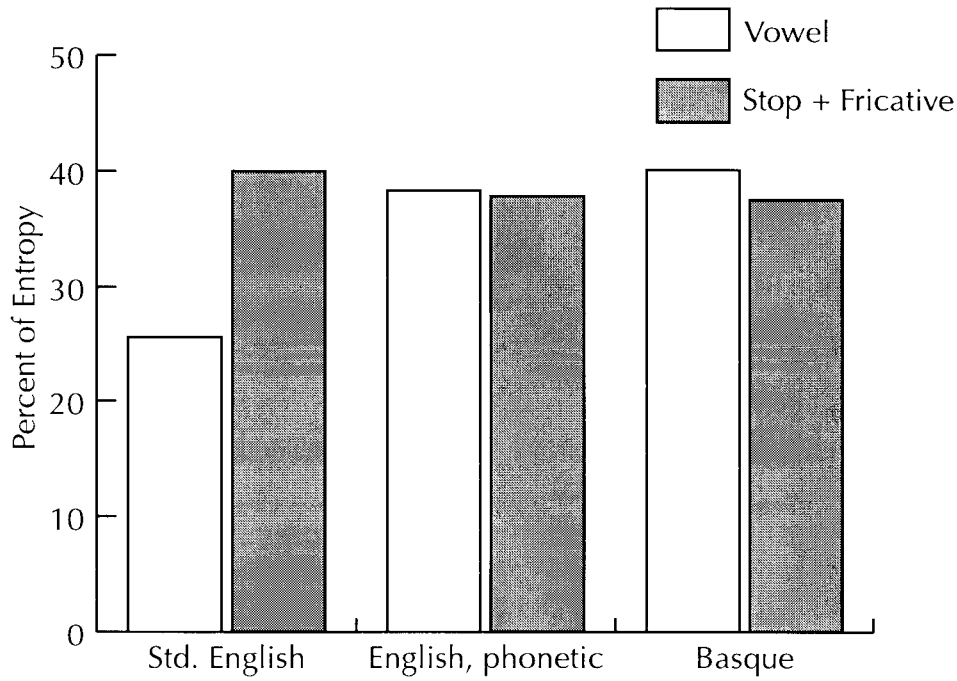


Figure 4.1: Percent of entropy in vowels and consonants for standard English spelling, phonetic transcription of English, and for Basque. Categories do not total 100% because vowel-like consonants have been excluded. □

4.3 Previous Work on Vowel Versus Consonant Recognition

Occasionally, though rarely, researchers will tabulate recognition results separately for vowels and consonants (Shoup 1980).

4.3.1 Conventional Recognition Architectures

The Sphinx speech recognition system (Lee et al. 1990), which used discrete-emission biphone HMMs (chap. 1), had higher recognition rates for stops than sonorants (table 4.4) (Lee and Hon 1988). The latter category includes vowels as well as certain consonants such as liquids, which in the present project were left in the "other" category.

4.3.2 Nonstandard Recognition Architectures

Nonstandard architectures may have greater problems with vowels than consonants. For example, a recognizer designed to mimic normal and impaired human hearing was tested and compared to

Sonorant	66%
Stop	70%
Fricative	78%
Closure	93%

Table 4.4: Previous work examining the recognition rates of vowels and consonants (Lee and Hon 1988). These categories differ from those used in the present project, since the vowels are not separated out from the other sonorants. The recognition system used context-dependent discrete-emission biphone HMMs and the cepstrum with delta and energy terms. □

human listeners (Giguere et al. 1997). That system's relative performance on vowels was worse than predicted by the abilities of the human subjects.

For this project, the Recnet recurrent neural network was downloaded and tested on the standard Texas Instruments–MIT speech database (TIMIT) (Garofolo et al. 1993), in order to tabulate its errors by phonetic class. Its overall phone error rate, which had previously been reported (Robinson 1994), was here confirmed to be 26.3%—coincidentally the same as ViaVoice's performance on the Wisconsin data (section 4.4). Examples from the literature of other phone-recognition error rates on TIMIT are 26.6% (Chang and Glass 1997) and 28.3% (Ström 1997).

4.3.3 Prior Work on Text Analysis

An early study on the statistics of spoken English (Denes 1963) tabulated minimal pairs of phonemes for words that differ by a single sound; for example, the "p" and "k" of "peep" and "keep." The study used phonetic transcriptions from books that teach English as a second language. There was no direct comparison of the importance of vowels and consonants, but for consonants the most common distinctions were based on manner (for example, "d" versus "z") of articulation rather than voicing (e.g., "f"/"v") or place of articulation. The above finding suggests that the microbeam data would be especially useful for vowel recognition, since the data directly encode place of articulation. Voicing would be more easily extracted from sound or measurement of laryngeal vibration. Manner could be determined from the dynamics of the microbeam data (as opposed to place, which can be determined on a frame-by-frame basis).

Phoneme Category	Acoustic-only Error Rate	Error Rate for Acoustic and Articulatory Input	Relative Improvement
Vowel	31.4%	23.6%	24.9%
Stop Consonant	35.1	32.6	7.1
Fricative	23.1	18.4	20.6

Table 4.5: Error rate reductions for stop and fricative consonants and vowels, for recognizer trained and tested with Wisconsin X-ray microbeam data □

4.3.4 Human Listeners

A related project at the Oregon Graduate Institute attempted to determine whether vowels or consonants were more important to comprehension (Cole et al. 1996). Those experimenters modified sentences from TIMIT (Garofolo et al. 1993) by replacing either vowels or consonants with noise; human subjects were then asked to transcribe the sentences. Due to coarticulation, some portion of the vowel sound may be used for identifying the neighboring consonant (Strange and Verbrugge 1976). The listeners' baseline performance on the unaltered sentences was 94% of words correctly recognized. Removal of consonant-like intervals reduced the word recognition rate to 57%, while removal of vowel-like intervals reduced the rate to 14%. Because a sharp delineation of consonant and vowel time intervals is not possible, the experiment is a little difficult to interpret.

4.4 Vowel and Consonant Recognition with Sound and Articulation

Three baseline continuous-speech recognizers were used. Two were conventional cepstrum-and-HMM systems: the acoustic-only recognizer of chapter 3, and a commercial large-vocabulary continuous speech recognizer, IBM ViaVoice Gold. These conventional systems were tested on speaker JW45 of the Wisconsin microbeam data set. A third system, Recnet, was a recurrent artificial neural network publicly available for download (Robinson 1994). Recnet was trained and tested using disjoint subsets of TIMIT.

The belief that vowels are better recognized seems based on the apparent compatibility of the cepstrum with the source-filter model of speech.

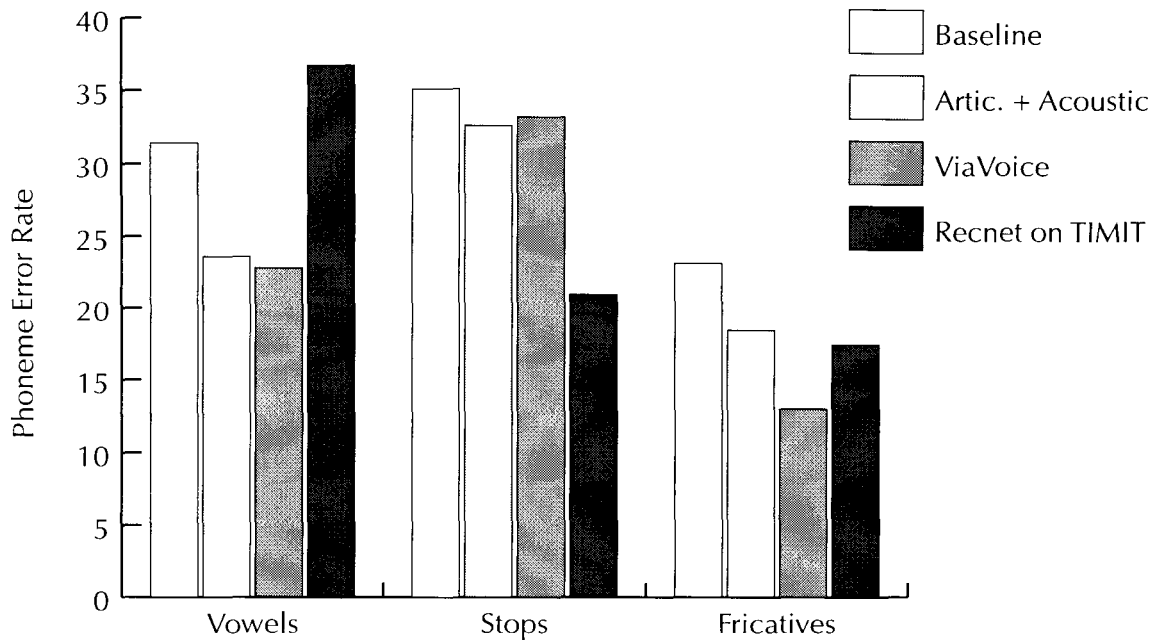


Figure 4.2: Phone error rates for vowels, stop consonants, and fricatives, for four recognizers: the acoustic-only baseline and acoustic-and-articulatory recognizer (both of chapter 3), IBM's ViaVoice Gold, and the Recnet recurrent neural network. The test data were six paragraphs spoken by subject JW45 in the Wisconsin microbeam data. Vowel recognition improved more than consonant recognition when articulation was added, and ViaVoice and Recnet performed at least as well on consonants (stops and fricatives averaged) as on vowels. □

4.5 Vowel and Consonant Recognition by the Commercially Available ViaVoice Recognizer

In another experiment, paragraph recordings (appendix F) from the Wisconsin x-ray microbeam data set (section 3.4.1) were presented to the commercially available IBM ViaVoice recognizer. Phone sequences were inferred from ViaVoice's text output as described in section 4.7.1.

Category	Phone Error Rate
Vowel	36.7%
Stop Consonant	20.9%
Fricative	17.4%
Overall	26.2%

Table 4.6: Recnet's (Robinson 1994) phone error rates for vowels, stops, and fricatives. □

4.6 Vowel and Consonant Recognition by Recnet Recurrent Neural Network

Recnet, a recurrent artificial neural network, was downloaded and tested on the TIMIT recordings (Garofolo et al. 1993). Although the overall phone error rate on TIMIT had been published by Recnet's creator (Robinson 1994), it was not tabulated separately for vowels and consonants. Unlike ViaVoice or the recognizer of chapter 3, Recnet performed far better on stops and fricatives than on vowels (table 4.6)—but since Recnet did not use a dictionary to constrain phone choice, the comparison is not straightforward.

4.7 Vowel and Consonant Recognition on Switchboard Telephone Conversations

The final stage of this work concerned errors made by a state-of-the-art recognizer transcribing telephone conversations. The goal was to see whether the results of section 4.3.1 held for a newer (by nine years) large-vocabulary recognizer.

Recognizer development and testing had been performed previously by other researchers, in the 1997 Large-Vocabulary Continuous-Speech Recognition Workshop at Johns Hopkins University's Center for Language and Speech Processing (Jelinek 1998a). The baseline recognizer's text output for the test data was provided by Joe Picone and Aravind Ganapathiraju of the University of Mississippi.

The present analysis started with transcriptions of telephone conversations from the Switchboard

Vowel	aa, ae, ah, ao, aw, ax, ay eh, ey, ih, iy, ow, oy, uh, uw
Stop	b, d, g, k, p, t
Fricative	dh, f, s, sh, th, z, zh
Affricate	ch, jh
Nasal	m, n, ng
Other	el, en, er, hh, l, r, sil, w, y

Table 4.7: Switchboard phonetic symbols grouped into vowels, stops, fricatives, affricates, and nasals □

data set (Godfrey et al. 1992). Since the transcriptions were composed of words, rather than phone-like units, additional processing was required to classify errors as vowel or consonant.

4.7.1 Determination of Phonetic Errors from Word-Level Transcripts

In general, a specific word sequence could correspond to multiple possible phonetic sequences. For each utterance in the present project, a pair of phonetic sequences was chosen—one for the recognizer's output, one for the hand transcription—having the minimum number of insertion and deletion errors (chap. 1). The procedure implemented was essentially the level-building algorithm (Myers and Rabiner 1981) (Rabiner and Juang 1993), but applied here to text-to-text correction instead of text-to-acoustic matching.

In the present version of level building, for each word, alternate pronunciations were determined using the standard dictionary from the workshop (Jelinek 1998a). These pronunciations used a similar phone set to ARPABET (Garofolo et al. 1993) and the Wisconsin phonetic transcription (section 4.2.1).

After the optimal pair of phonetic sequences was determined, deletion errors—phone present but missed by recognizer—were tabulated as vowel, stop, fricative, affricate, nasal, or other. Insertion errors were ignored, whether or not they were part of a substitution error; each substitution error was considered to be an insertion plus a deletion. The weights for substitution and deletion errors were therefore equal. A list of the phones in each category appears in table 4.7. The larger size of the present transcripts compared to those of section 4.2.1 gave an adequate sample size for more and finer categories.

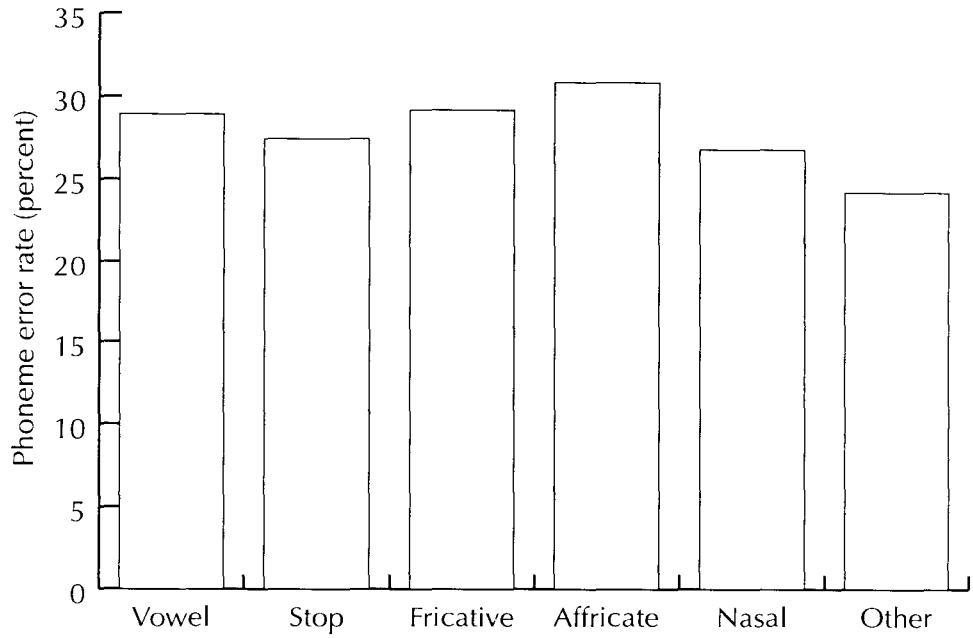


Figure 4.3: Phoneme error rates on telephone conversations in the Switchboard data set, when transcribed by a state-of-the-art large-vocabulary recognizer □

4.7.2 Vowel and Consonant Recognition: Switchboard Results

The recognizer's errors on the Switchboard telephone conversations are broken down by phone class in figure 4.3. For the three main categories—vowel, stop, and fricative—errors are not dramatically different. Stops are recognized slightly better than vowels, and recognition is best on the default "other" category, which is fundamentally different because it includes pauses. An affricate (e.g., "ch") can be thought of as a stop ("t") followed by a fricative ("sh"), requiring two successful identifications in a row; this helps explain why the error rate was higher for affricates than stops or fricatives.

4.7.3 Consequences of the Grammar Model and Dictionary

The Switchboard results were based on recognition using a grammar model and dictionary. Therefore, the variation in error rate across phone classes does not represent the relative difficulty of purely acoustic recognition. Instead, it represents the difficulty of recognizing the different sounds given a higher-level model of speech. In principle it is possible to test recognition with a triphone-loop grammar, but the original triphone models were not available for this project—only recognizer

output. It might also be possible to extend an existing theory for the interaction of acoustic and grammar models (Ferretti et al. 1990) to isolate the acoustic problem.

4.8 Recognition and Entropy of Vowels and Consonants: Conclusions

State-of-the-art recognizers are not necessarily better at recognizing vowels than consonants. Further, when only stop and fricative consonants are considered against the vowels, and a phonetic transcript is used, the entropy of consonants and vowels is comparable. This suggests that, since consonants are not the current weak point of recognizers, it may not be necessary to concentrate research efforts on them in order to obtain a dramatic recognition improvement. These two results call into question the vowel-consonant discrepancy identified in a standard text (see quote at head of this chapter) (Rabiner and Juang 1993).

4.9 Future Work

Both information content and recognition performance for vowels and consonants can be better evaluated. Information-content estimates can be improved by entropy calculation, reading experiments with humans, or presentation of modified speech. Grammar model effects can be factored out of the recognizer, and nonstandard recognition architectures can be assessed.

4.9.1 Calculation of Entropy from Text

The calculated estimates of information conveyed by vowels and consonants would benefit from a more sophisticated predictive model (Brown et al. 1992), or at least the use of second-order transition statistics.

4.9.2 Textual Experiments with Human Listeners

Since humans outperform the best algorithms at predicting what letter will follow an initial sequence, the question of vowel versus consonant entropy in text should be tested with human listeners. Gambling techniques have been successful for calculating overall entropy (Cover and King 1978), and should be readily adaptable to separate tabulation of vowels and consonants.

4.9.3 Audio Experiments with Human Listeners

Previous studies of speech intelligibility with vowels or consonants removed (Cole et al. 1996) could certainly be removed. In particular, the problem of listeners being able to identify consonants from coarticulation with vowels (Strange and Verbrugge 1976) can be addressed by using synthetic speech instead of recordings of people.

Some synthesis parameters could be derived from recordings, to give a more natural sound. However, the formant frequencies should be modified to eliminate transition cues: instead of deleting intervals, the targeted phones can be replaced by canonical sounds (e.g., /ə/ for all vowels). The surrounding phones would then coarticulate with the new, neutral phone, instead of retaining information about the missing one.

4.9.4 Experiments with Other Recognition Architectures

Speech recognizers have been built on rather different principles, such as neural networks (Bourlard et al. 1994). It is not clear that these other architectures have the same strengths and weaknesses as the standard cepstrum-and-HMM approach, so any general conclusions about vowel and consonant recognition should be tested on them as well.

4.9.5 Analysis of Recognizer Performance

As described above (section 4.7.3), it may be possible to factor out the effect of the grammar model on vowel and consonant recognition.

Chapter 5

Discussion and Conclusions

In this thesis, to improve speech recognition performance, audio signals were combined with different types of auxiliary kinematic information. Direct measurements of articulator position—x-ray microbeam data—improved continuous speech recognition accuracy (chap. 3). Preliminary results suggest that side-view lipreading with audio and video may outperform recognition with audio alone (chap. 2). A hypothesized shortcoming of existing systems—consonant recognition—was found to be less of a problem than previously supposed, both for the new recognizer presented here and for conventional state-of-the-art systems (based either on the cepstrum and HMMs or on recurrent neural networks). It is not addressed whether consonants were a weak point in old approaches to recognition, nor whether they will present problems for future recognizers. In the approach described in this thesis, kinematic information did not improve consonant recognition as much as vowel (chap. 4).

The following sections address advantages and disadvantages of side-view lipreading and motor representations for speech recognition. The latter part of this chapter also discusses the following questions:

How many articulatory degrees of freedom are important?

Are state-of-the-art (cepstrum-and-HMM and recurrent neural network) speech recognizers less accurate at recognizing consonants than vowels, and do consonants convey more information?

5.1 Advantages and Disadvantages of the Side-View Approach to Lipreading

The preliminary recognition results in this thesis suggest advantages to using of side-view lipreading (section 2.8). For example, this approach requires far less computation than conventional lipreaders using dynamic contours (snakes). Another advantage is that it is stateless, so tracking errors do not propagate from one frame to the next. An analysis of errors made by the audio and video classifiers described in chapter 2 suggests that the two modalities are complementary (section 2.8.3). Combining audio and video reduced errors compared to either alone; this property of independence of the channels is true of front-view lipreading as well. The accuracy of the two approaches on the same example utterances has yet to be compared. The confusion matrix in table 2.2 suggests that the side-view video pipeline readily distinguishes important features of articulation.

The side-view processing pipeline consists of a median filter followed by a threshold, state machine, and centroid calculation. Dynamic contours (snakes) involve blurring the image, using edge information to make forces along a contour, and a dynamic simulation of the snake's motion (Kass et al. 1987) (Platt 1989). If the image changes in an unexpected way or too quickly, the snake may fail to track the intended edge; such tracking errors propagate from frame to frame.

A disadvantage to the approach reported here is that it uses a camera worn on the head—currently in a mask, but perhaps in the future on a headest instead. Even for a modern lightweight camera on a single circuit board, this may be too obtrusive for certain applications.

5.2 Advantages and Disadvantages of Motor Representations in Speech Recognition

Another set of issues addressed in this thesis are the following:

- Can recognition with joint acoustic-articulatory representations outperform recognition with acoustics alone?

The above is closely related to the following question regarding the motor theory (section 1.2.2):

- Are articulatory representations more invariant than acoustics, and therefore a good choice of an intermediate domain in speech recognition?

Using a joint acoustic-articulatory feature set, as in this thesis, hypothesizes a combination of the two strategies for perception.

Advantages

A statistically significant error rate improvement—7.4%—was obtained by appending articulation to acoustics (table 3.6). The improvement from adding articulation was greater for vowels and fricatives than for stops (section 5.4), and for male speakers compared to female. A hypothesis about the discrepancy between the sexes is that because females' vocal tracts are smaller, the variability between different instances of a phone is large compared to the distance between different phone categories.

Thus, the strategy of combining articulatory with acoustic representations outperformed the conventional acoustic-only, cepstrum-and-HMM approach.

Here, with only about 10 minutes of training data per speaker, a statistically significant improvement was seen. If motions can be recovered from sound, large data sets such as the Switchboard corpus (section B.7) can be augmented with recovered articulatory representations. Training the recognizer with such data can be expected to further improve recognition.

Disadvantages

A negative aspect of the present results is that recovered motions will never be as exact as the actual measurements used here. Another unforeseen and unfortunate result is that articulation aided recognition of male speakers far more than female.

The use of principal components analysis (PCA) in chapter 3 facilitates interpretation (section 5.3) and implementation. However, recognition may be more accurate with a different parameterization of motion.

5.3 How Many Articulatory Degrees of Freedom Are Important?

If, as described above, articulatory representations might aid recognition, how many degrees of freedom of the articulators are relevant to recognition? In this thesis, recognizer performance after the first round of training (flat start; section 3.9.4) was the same for 4, 6, or 8 articulatory features (section 3.12.2). The first 4 features describe 90.2% of the variation in lip, tongue, and jaw position (table A.1), and qualitatively seem to capture a number of phonetic contrasts (section A.3.3).

The above results suggest that relatively few articulatory parameters may be sufficient. It is hard to conclusively determine the number of degrees of freedom because of the following considerations:

1. Additional information—in the sense of independent measurements—never reduces the discrimination ability of an ideal classifier. At worst, the new information might be irrelevant and discarded. At best, it might enable distinctions which were not possible from previous measurements. In an intermediate case, it could simply be used to reduce measurement noise. The corollary to this is that reducing the number of degrees of freedom would not increase the performance of an ideal classifier. These observations are strictly from an information-theoretic perspective. For a non-ideal classifier, different representations of the same information (for example, articulation recovered from sound versus the sound itself) can result in different levels of performance.

2. What types of parameterizations are considered? Linear transformations of coordinates were used here, but nonlinear mappings could also be very useful.

3. In what context, and with what data set, is classification ability to be measured?

4. What level of performance is acceptable? Recognition accuracy comparable to humans would allow one to start making general claims. Such performance is probably a long way off.

5.4 Are State-of-the-Art (Cepstrum-and-HMM and Recurrent Neural Network) Speech Recognizers Less Accurate at Recognizing Consonants than Vowels, and Do Consonants Convey More Information?

It makes sense to concentrate speech-recognition research on the weak links of existing systems. Are consonants currently a problem—harder to recognize and more important than vowels? The results of chapter 4 suggest not. In that chapter, entropy of text was calculated separately for vowels, stop consonants, and fricatives. This showed that with a phonetic transcription, the latter two together conveyed about as much information as vowels. For the different phone classes, errors made by the following recognizers were tabulated:

- A conventional state-of-the-art recognizer, the baseline system used in the 1997 Large-Vocabulary Continuous-Speech Recognition Workshop at Johns Hopkins University;
- The joint acoustic-articulatory recognizer in chapter 3.
- The acoustic-only recognizer of chapter 3.
- A commercially available recognizer, IBM's ViaVoice Gold; and
- Recent, a recurrent neural network.

In all cases, the average error rates for stop consonants and fricative were at least as low as the error rates for vowels.

Among the consonants, only stop consonants and fricatives were considered here, since they are the least vowel-like. Phonetic transcriptions are more relevant than English orthography because during recognition, words are treated as single symbols and mapped to sequences of phonetic symbols, not letters.

Previous work suggested that people use vision to determine manner of articulation (e.g. /t/ versus /s/) and audition to determine place (e.g., /k/ versus /t/) (MacDonald and McGurk 1978). Another researcher found that, in the statistics of spoken English, manner was more important than place for consonant contrasts (Denes 1963). X-ray microbeam measurements, which directly measure place of articulation, can therefore be expected to improve vowel recognition more than consonant; this prediction held true for the experiments of this thesis (section 4.4).

Appendix A

Principal Components Analysis of Microbeam Data

A.1 Definitions: PCA and MLG

A.1.1 Principal Components Analysis (PCA)

Given a sample set of vectors, the principal components are directions in which the sample set has its greatest variance. To be precise, the principal components are an orthogonal, ordered set of eigenvectors of the sample set's covariance matrix. If the sample set is represented as $\mathbf{x}(t)$ for integers $1 \leq t \leq N$, the entries $\mathbf{C}_{ij}$ of the covariance matrix are

$$\mathbf{C}_{ij} = \frac{1}{N-1}\left[\sum_{t=1}^{N}\left(\mathbf{x}_i(t) - \frac{\sum_{\tau=1}^{N}\mathbf{x}_i(\tau)}{N}\right)\left(\mathbf{x}_j(t) - \frac{\sum_{\tau=1}^{N}\mathbf{x}_j(\tau)}{N}\right)\right]$$

The principal components $\mathbf{p}_k$ satisfy the following characteristic equation:

$$\mathbf{C}\mathbf{p}_k = \lambda_k \mathbf{p}_k$$

Each principal component describes a fraction of the total variance of the sample set.

A.1.2 Maximum-Likelihood Gaussian (MLG) Classifier

A maximum-likelihood Gaussian (MLG) classifier is similar to the continuous emission density HMMs which are used in speech recognition. Specifically, an MLG classifier is equivalent to a single-state HMM with a single Gaussian in its emission density. The following discussion describes MLGs for simplicity but these considerations apply to HMMs as well. For an input x, the classifier generates a set of probability scores $p_i(x)$ for the categories $1 \leq i \leq N_c$. These scores are determined by the category means and variances. For one-dimensional inputs, the scores are computed as follows (Duda and Hart 1973):

$$p_i(x) = \frac{1}{\sqrt{2\pi}\sigma_i} e^{-(x-\mu_i)^2/(2\sigma_i^2)}$$

Equation A.1: Maximum-likelihood Gaussian classifier (MLG) for one-dimensional data: probability score $p_i(x)$ for category i with input x. □

The output of the MLG classifier, in response to an input x, is the category index i which maximizes $p_i(x)$. In practice, μ and σ above are estimated from training data. In the general case of HMMs, the statistics are iteratively estimated using expectation-maximization (E-M).

A.1.3 Single Transformation Matrix for All Categories

In the single-transformation-matrix approach, examples of all categories are pooled when computing the covariance matrix $\mathbf{C}$ defined in section A.1.1. A transformation matrix $\mathbf{A}$ is generated with N_D rows, each of which consists of one of the first N_D principal components. This matrix transforms all inputs before presentation to the MLG (section A.1.2). This reduces the dimensionality of the input data to N_D for recognition.

$$\hat{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t)$$

The means and variances for all categories are also projected into the subspace:

$$\hat{\mu}_i = \mathbf{A}\mu_i$$

$$\hat{\mathbf{C}}_i = \mathbf{A}\mathbf{C}_i\mathbf{A}^T$$

In the one-dimensional case, the above values are scalar. The category scores are obtained by substituting the transformed values into equation A.1:

$$p_i(\hat{x}) = \frac{1}{\sqrt{2\pi}\hat{\sigma}_i} e^{-(\hat{x}-\hat{\mu}_i)^2/(2\hat{\sigma}_i^2)}$$

A.1.4 Category-Dependent Transformation Matrices

Alternately, separate transformations $\mathbf{A}_i$ can be used for the categories i. In this case, the input is partitioned and only examples of a particular category are used to calculate that category's covariance matrix. Each input $\mathbf{x}(t)$ is then transformed N_c times to create intermediate representations $\mathbf{x}_i(t)$ for each category.

A.2 Previous Work on PCA of Speech Production Data

A.2.1 Sagittal Factor Analysis of Tongue (Harshman et al. 1977)

Other researchers have used factor analysis (closely related to PCA) to analyze sagittal tongue shape (i.e., in the front-back and up-down plane) (Harshman et al. 1977). That project was meant to provide an objective, data-driven basis for describing the tongue's degree of freedom. The analysis used cross-sectional width of the vocal tract at 18 locations, and the correlation (96%) between predicted and actual tongue configurations was used for assessment. Earlier projects had shown that two or three parameters were effective for describing the variations in tongue shape that were relevant to vowel production (Liljenkrants 1971).

A.2.2 Cross-Sectional Tongue Shapes for Vowels (Stone et al. 1997)

A previous study (Stone et al. 1997) applied PCA to ultrasound cross-section images of the tongue as vowels were spoken. Five examples each of 11 vowels in 2 consonant contexts were pooled for a global PCA; all data came from a single speaker. The first two principal components were found to account for 93% of the variance of the cross-sectional shapes, and defined a transformation projecting the data into only two dimensions. This represented a dramatic reduction from the seventy curve parameters initially extracted from each ultrasound image. After projection of all examples using this global (across phones) transformation, the researchers showed that front, back, and high vowels could still be distinguished.

The results presented in this appendix are discussed in section A.4 in the context of the above-cited study (Stone et al. 1997).

A.2.3 Across-Class PCA in Wisconsin X-Ray Microbeam Data (Roweis 1999)

The principal components of the tongue coordinates in the Wisconsin data have been previously calculated and plotted (Roweis 1999). That analysis showed the following when sagittal motion was considered separately from the other landmarks (beads): The first component, predominantly vertical, described tongue motion towards and away from the palate; the second, largely horizontal, described inward and outward motion of the tongue in the mouth; and the third described the tilt (front/back) of the tongue.

The present thesis reports the application of PCA jointly to all the landmarks in the Wisconsin data. The first four principal components, using a single analysis for all phones, appear in figure A.2 through figure A.5. The ordering of the components is different than found in the tongue-only previous work (Roweis 1999), but the first four components as a group seem qualitatively to describe the same tongue motions.

A.3 Results

A.3.1 Rationale for Using Single Transformation Matrix

Chapter 3 presents recognition results for microbeam coordinates transformed by a global PCA matrix. Advantages of the global transformation over multiple, phone-dependent transformation matrices include:

- Ultimately, it would be useful to know how many degrees of freedom of the articulators are necessary for classification. Although this thesis does not answer this general question, it does compare different numbers of degrees of freedom for a particular recognizer architecture. In the alternate approach—category-dependent input transformations—the input dimensionality is not reduced. For example, even if only one principal component were retained per category, the present project's use of 45 monophones would transform 16 position coordinates into 45 before recognition. Previous work has similarly used a single transformation matrix across phones, in order to understand the number of degrees of freedom of the articulators (Stone et al. 1997).

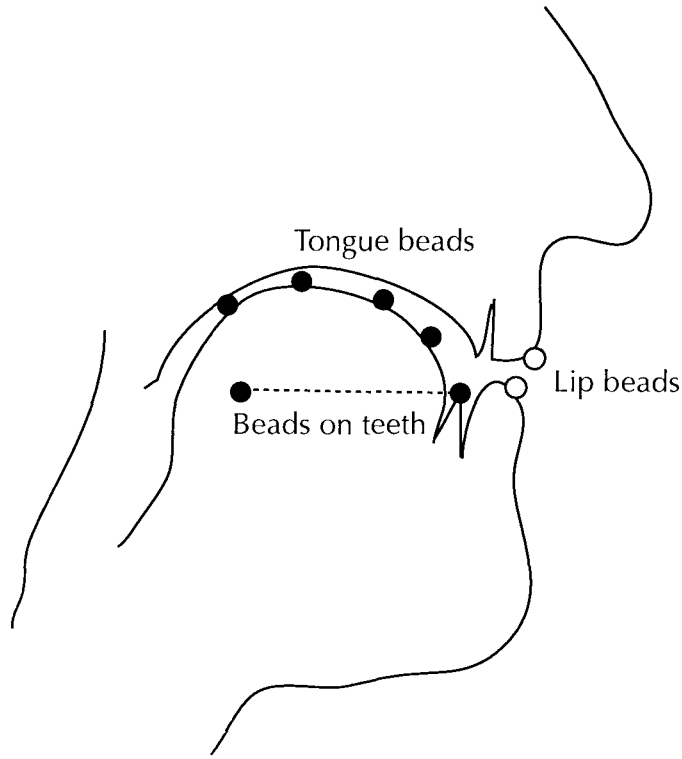


Figure A.1: Reference for figure A.2 through figure A.16: arrangement of beads on head for microbeam recording. This figure also appears as figure 3.3. □

- Fewer model parameters are required with a single transformation. With 45 categories, 16 input coordinates (as used here), and 4 principal components retained, the category-dependent matrices would have a total of 2,880 entries, versus 64 for the global transformation.

The disadvantage to implementing only a single transformation matrix is that multiple transformations could perhaps result in better performance.

A.3.2 Percent of Variance Explained by Principal Components

In figure 3.7, the percent of variance explained by the principal components is displayed for six speakers individually. This statistic pooled across speakers is shown in table A.1.

A.3.3 Statistics for Individual Phonetic Classes

The first principal component (direction of greatest variance) for each of eight phone classes and three speakers appears in figure A.14 and the following two figures. For each bead placed on the

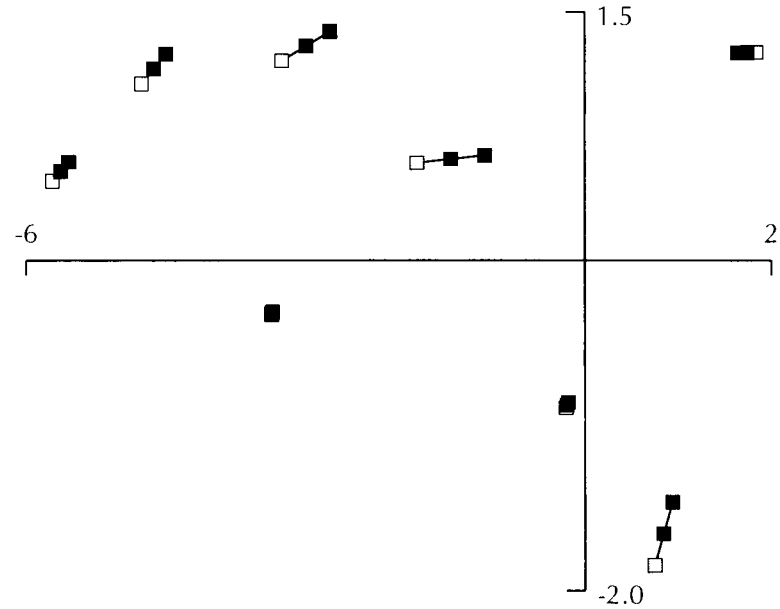


Figure A.2: First principal component for JW12, calculated with pooled data from all phones. See figure A.1 for a guide to landmark (bead) placement. □

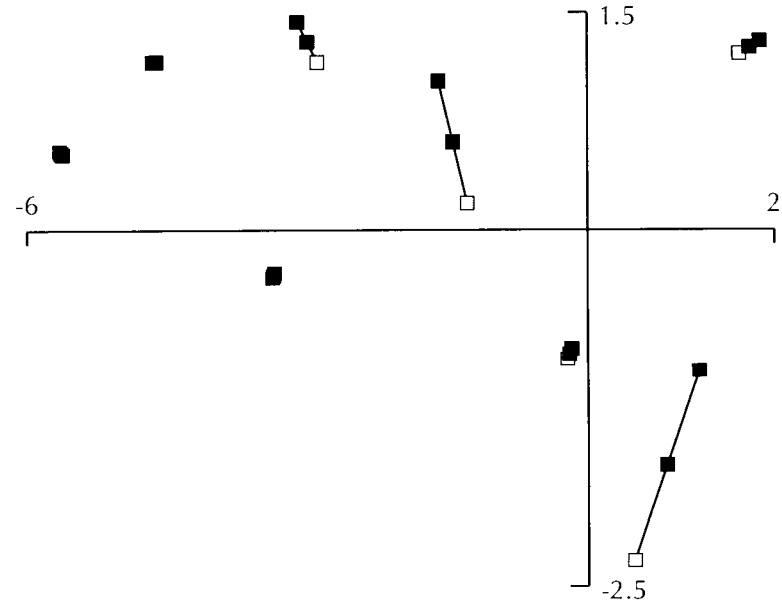


Figure A.3: Second principal component for JW12. □

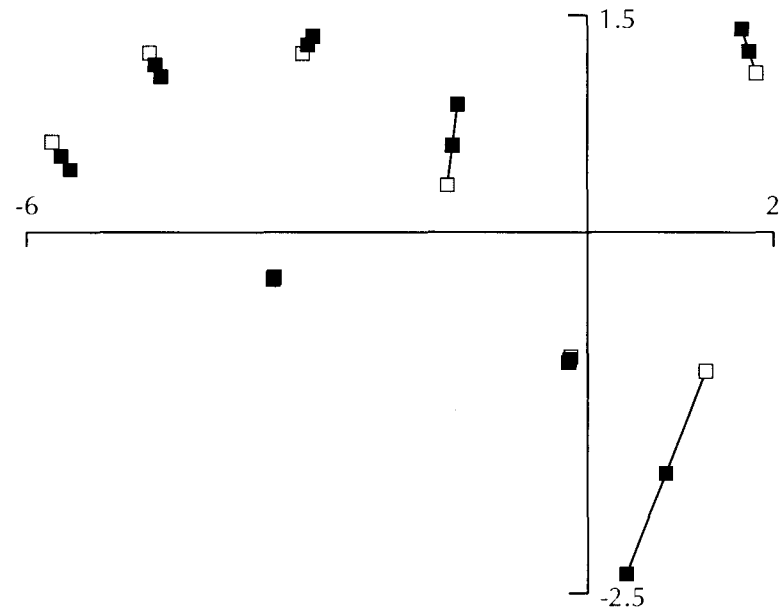


Figure A.4: Third principal component for JW12. □

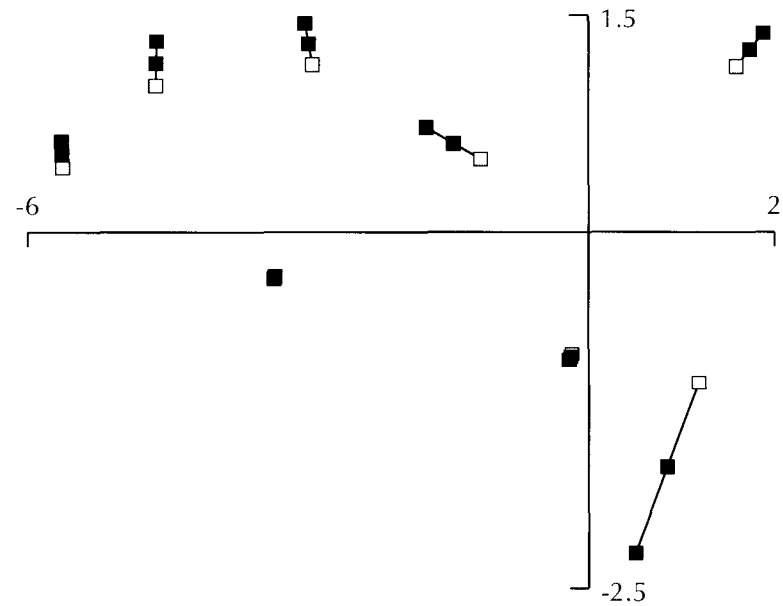


Figure A.5: Fourth principal component for JW12. □

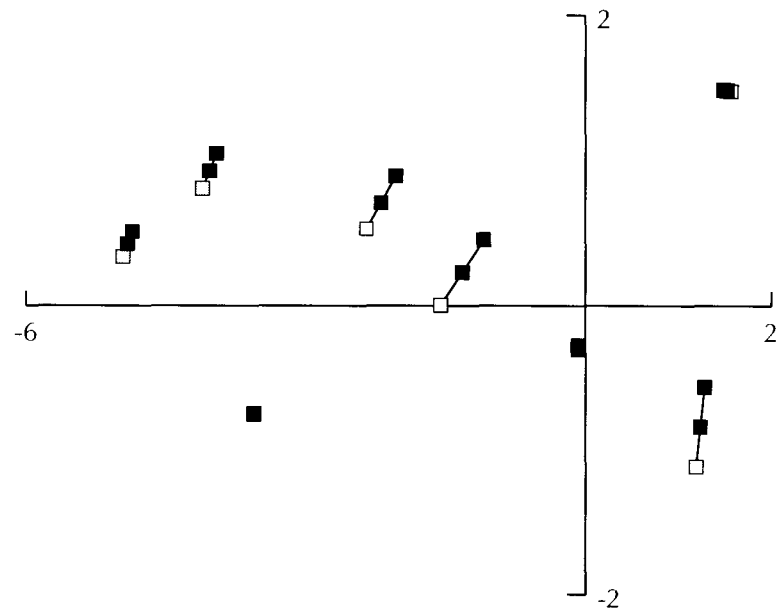


Figure A.6: First principal component for JW15, calculated with pooled data from all phones. □

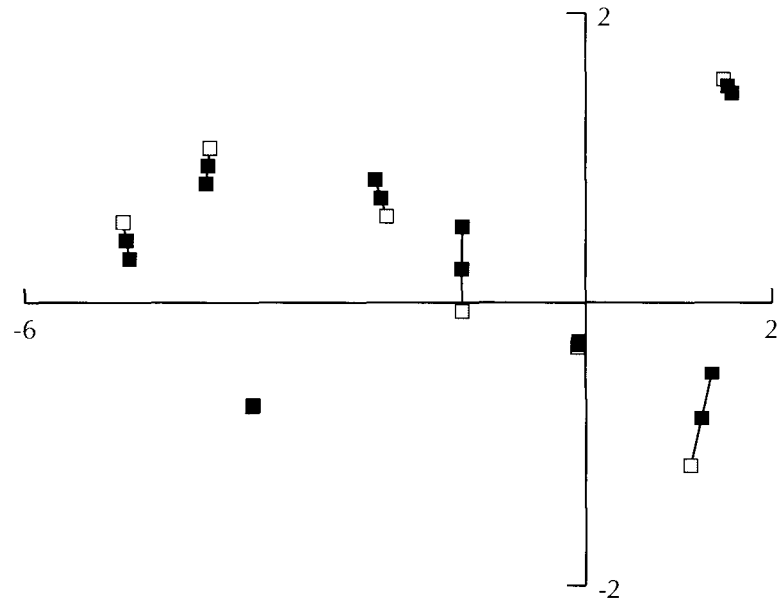


Figure A.7: Second principal component for JW15. □

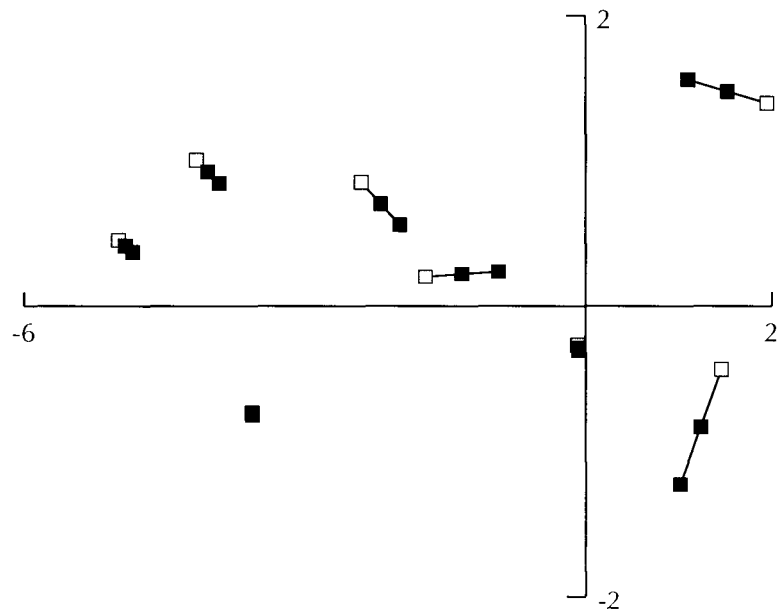


Figure A.8: Third principal component for JW15. □

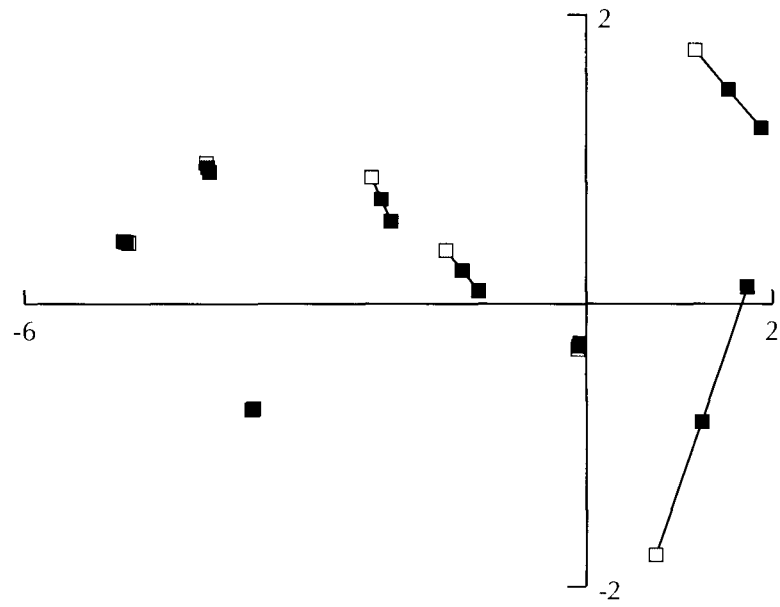


Figure A.9: Fourth principal component for JW15. □

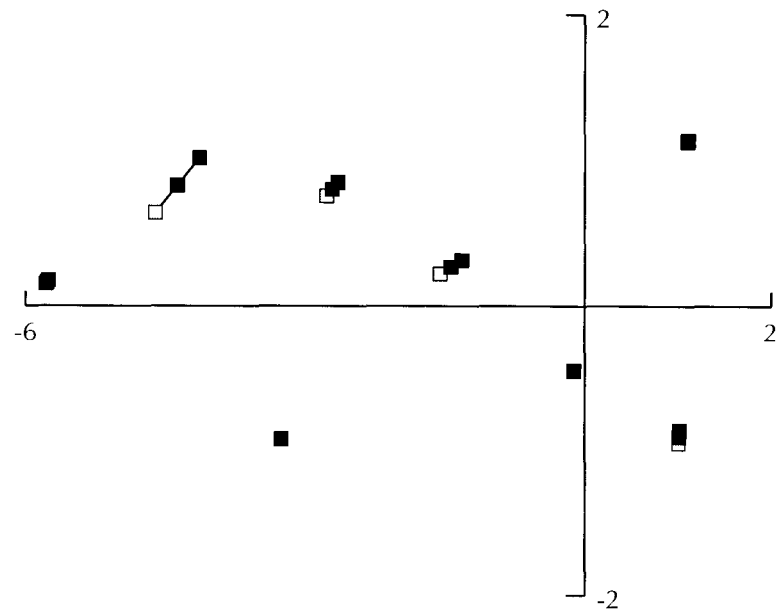


Figure A.10: First principal component for JW27, calculated with pooled data from all phones. □

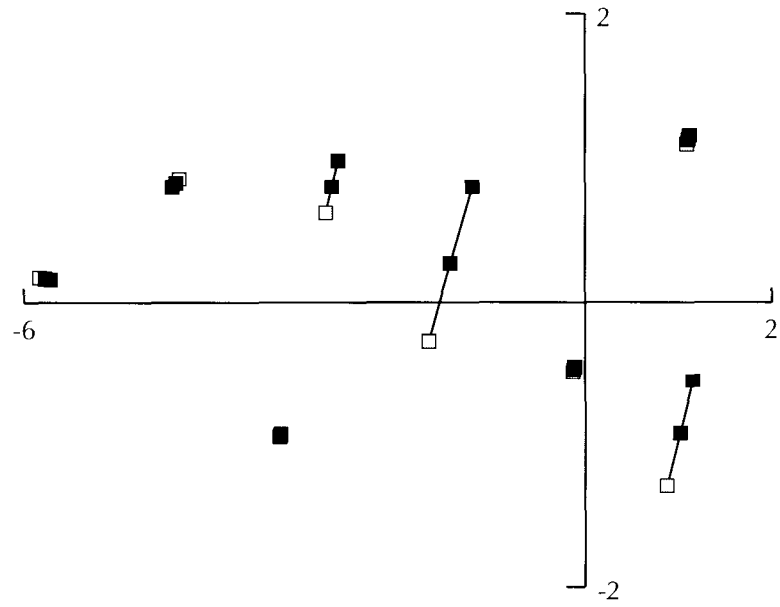


Figure A.11: Second principal component for JW27. □

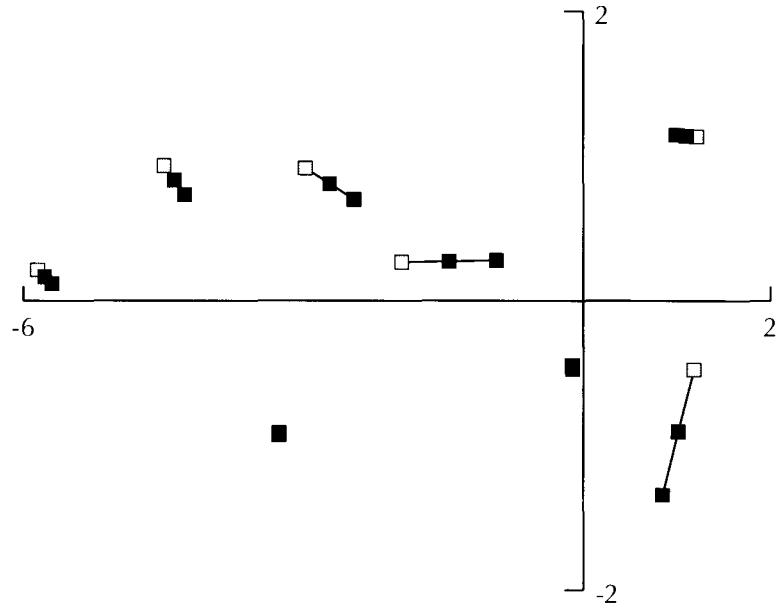


Figure A.12: Third principal component for JW27. □

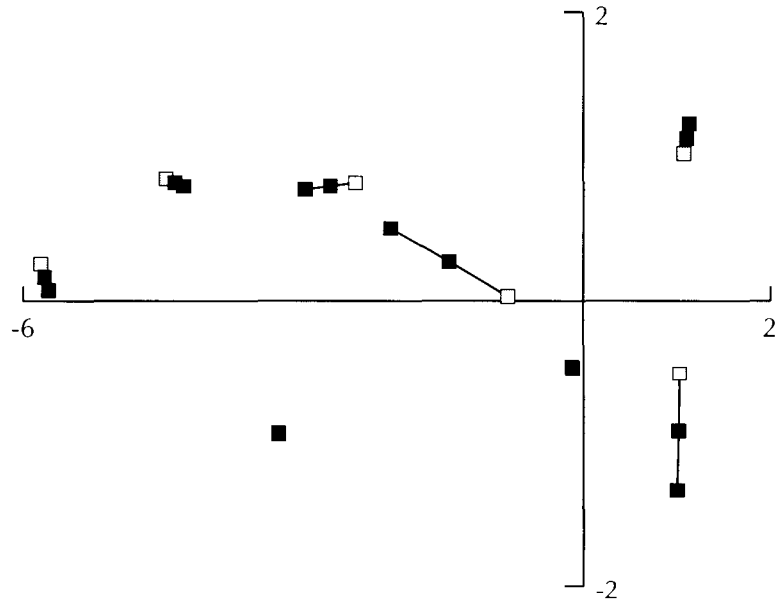


Figure A.13: Fourth principal component for JW27. □

n_{PC}	Variance explained
1	44.0%
2	66.7
3	80.3
4	90.2
5	93.6
6	95.9
7	97.2
8	98.1
9	98.7
10	99.1
11	99.5
12	99.7
13	99.8
14	99.9
15	100.0
16	100.0

Table A.1: Variance explained by first n_{PC} principal components, for all speakers combined. These components include lip, jaw, and tongue motion, so it is not surprising that more components (4) are required here to capture 90% of the variance than the 2 found in previous work on cross-sectional tongue shapes (Stone et al. 1997). □

mouth during x-ray microbeam data collection, a line segment is plotted extending three standard deviations on either side of the mean, in the direction of the first principal component. Note that this principal component represents variance within the phone class, not between one class and the rest of the data.

In figure A.17 and figure A.18, the first two within-phone principal components are projected by the single transformation matrix (per speaker) used for recognition in chapter 3. This transformation had a four-dimensional target space defined by the first four principal components of global, across-phone PCA.

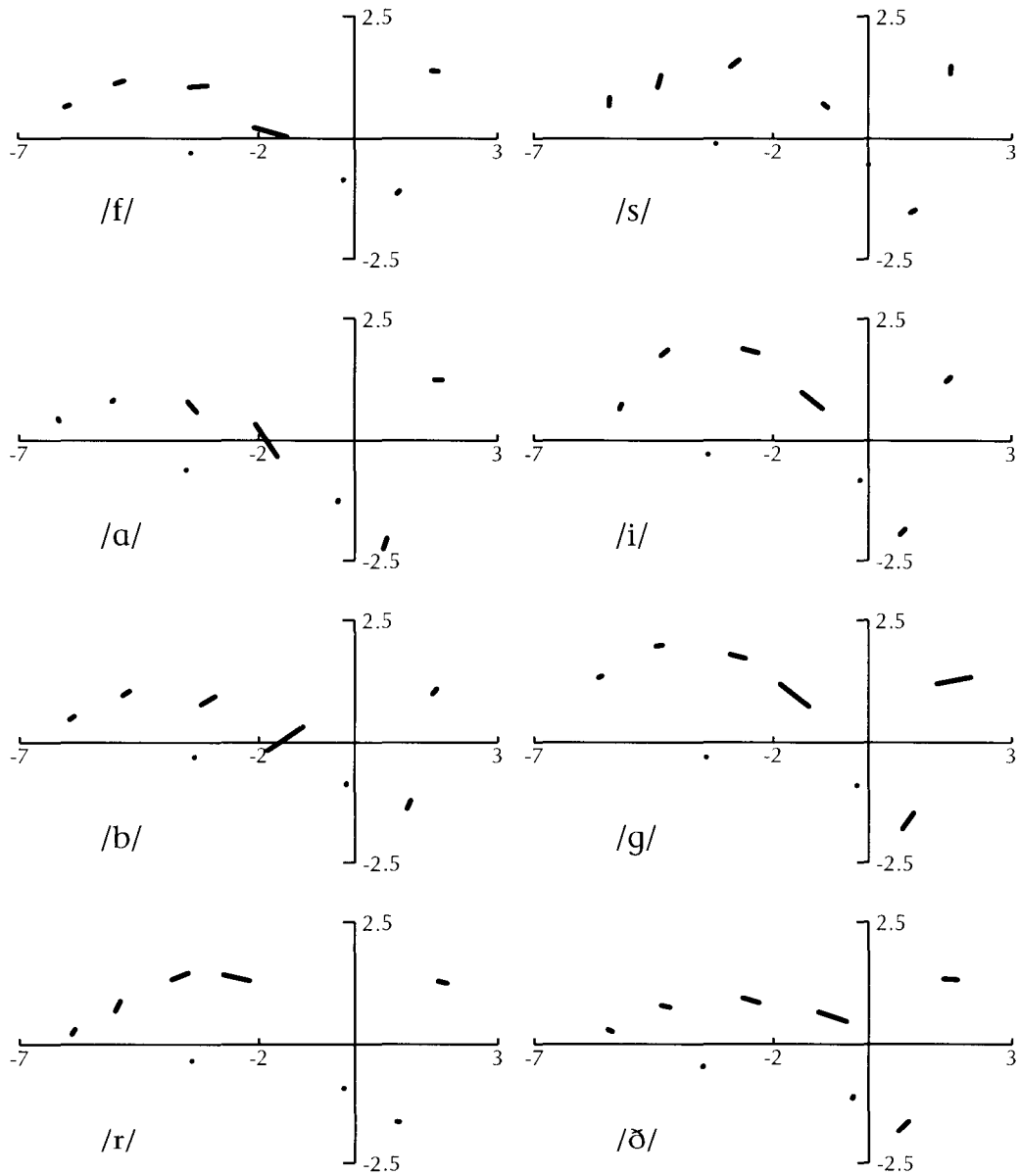


Figure A.14: First principal component for various phones as spoken by subject JW12. Line segments extend, in the same direction as the first principal component, for three standard deviations above and below the mean. See also table A.2. Non-critical articulators (e.g., the tongue for /f/) vary more than critical (lower lip for /f/), as observed in previous work (Papcun et al. 1992). □

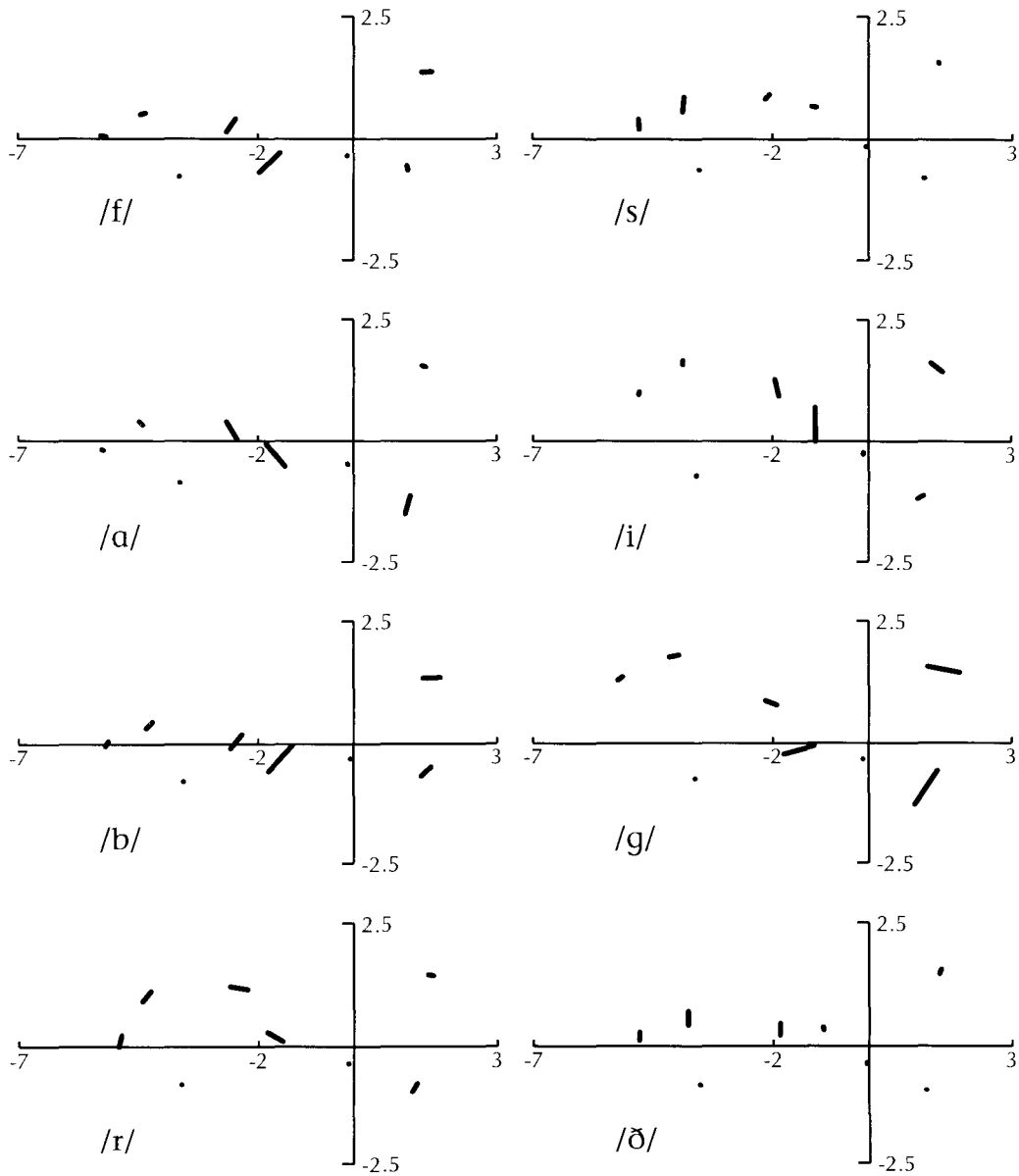


Figure A.15: Same as figure A.14 (first principal component for various phones), here plotted for speaker JW15. □

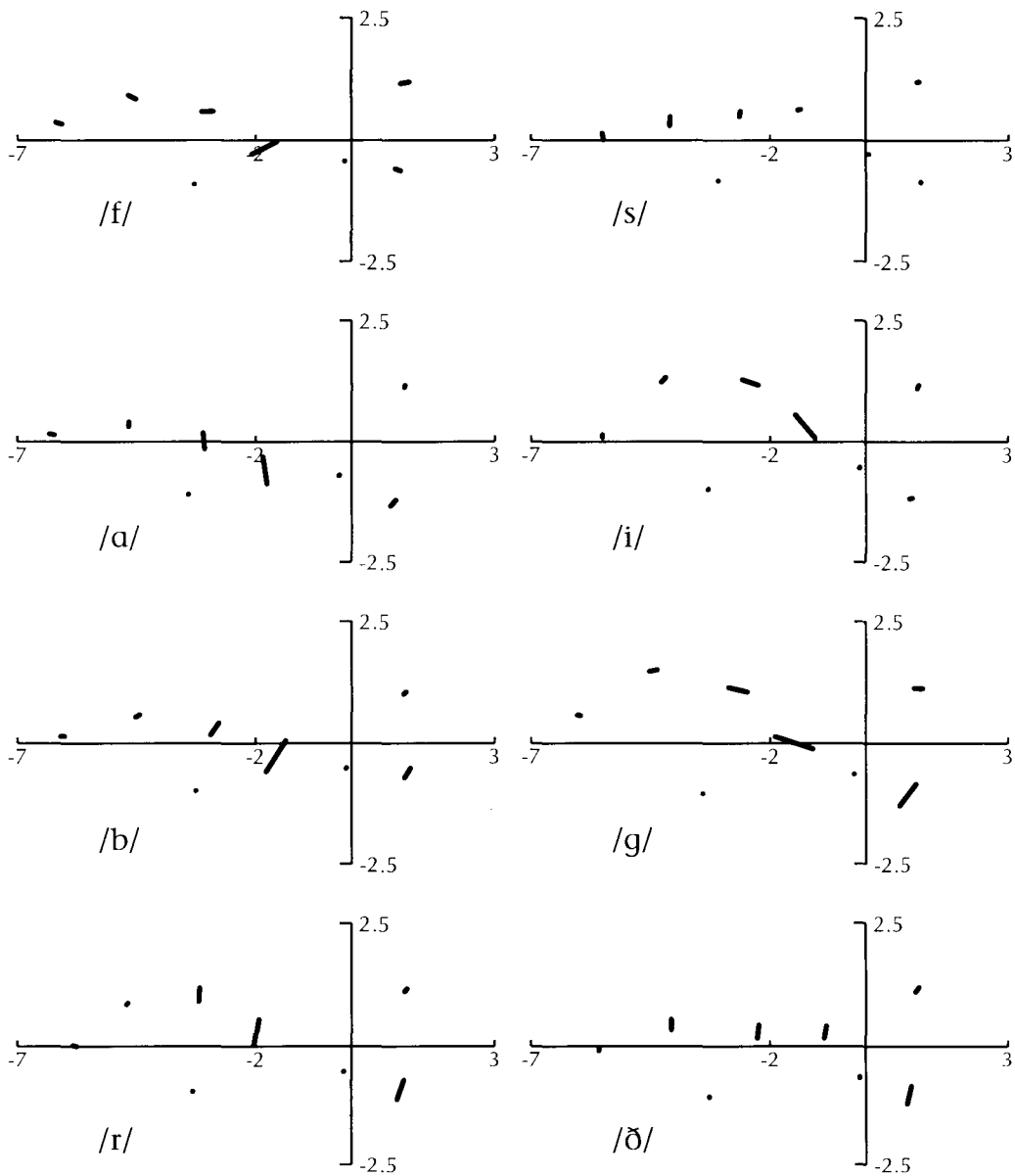


Figure A.16: Same as figure A.14 and figure A.15 (first principal component), but for speaker JW27. □

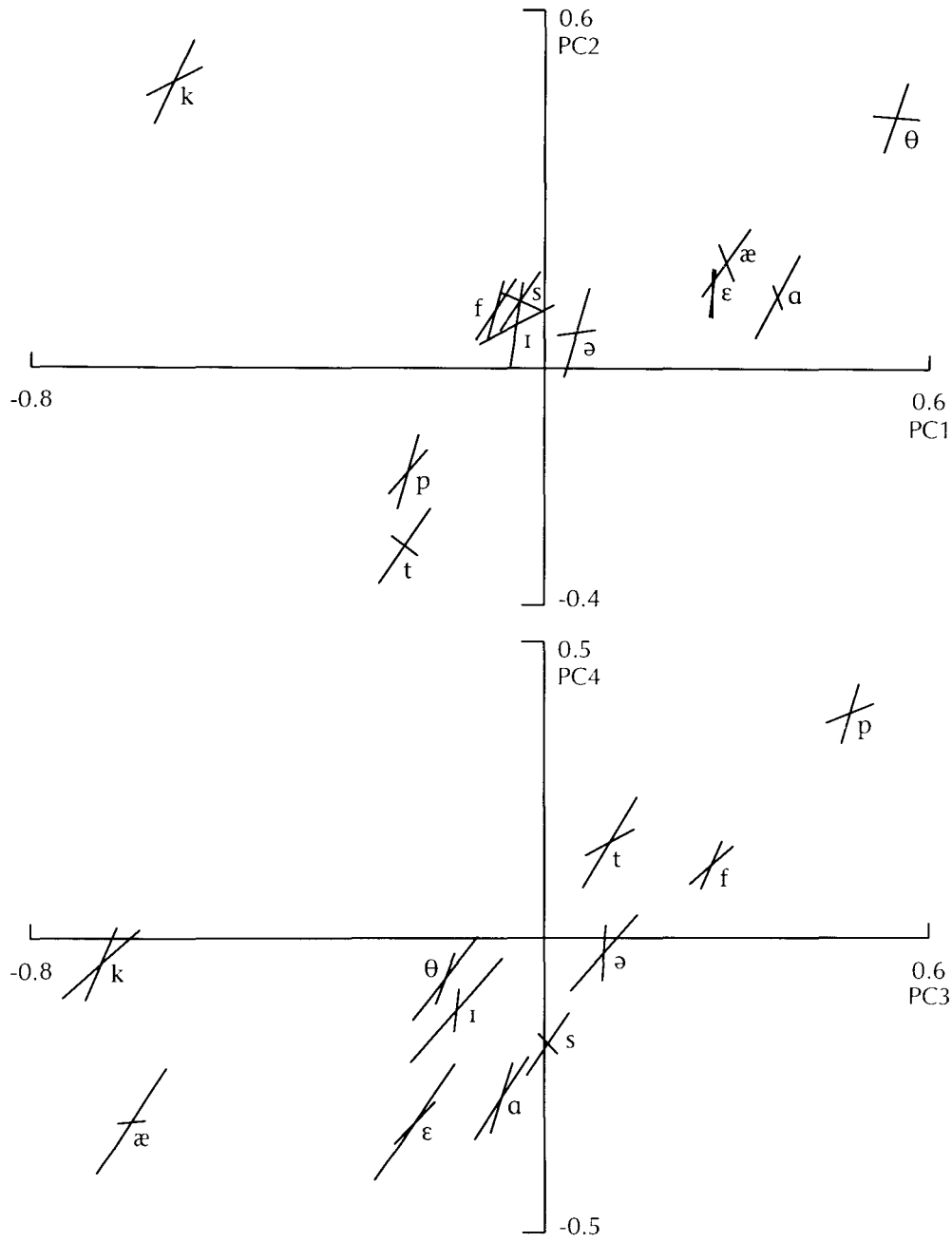


Figure A.17: Each phone's mean and first two principal components (PCs), projected onto the space of the first four global PCs. Each phone PC line segment extends for one standard deviation on either side of the mean The phones seem easier to distinguish using global PCs 3 and 4 than with PCs 1 and 2 (see discussion below). □

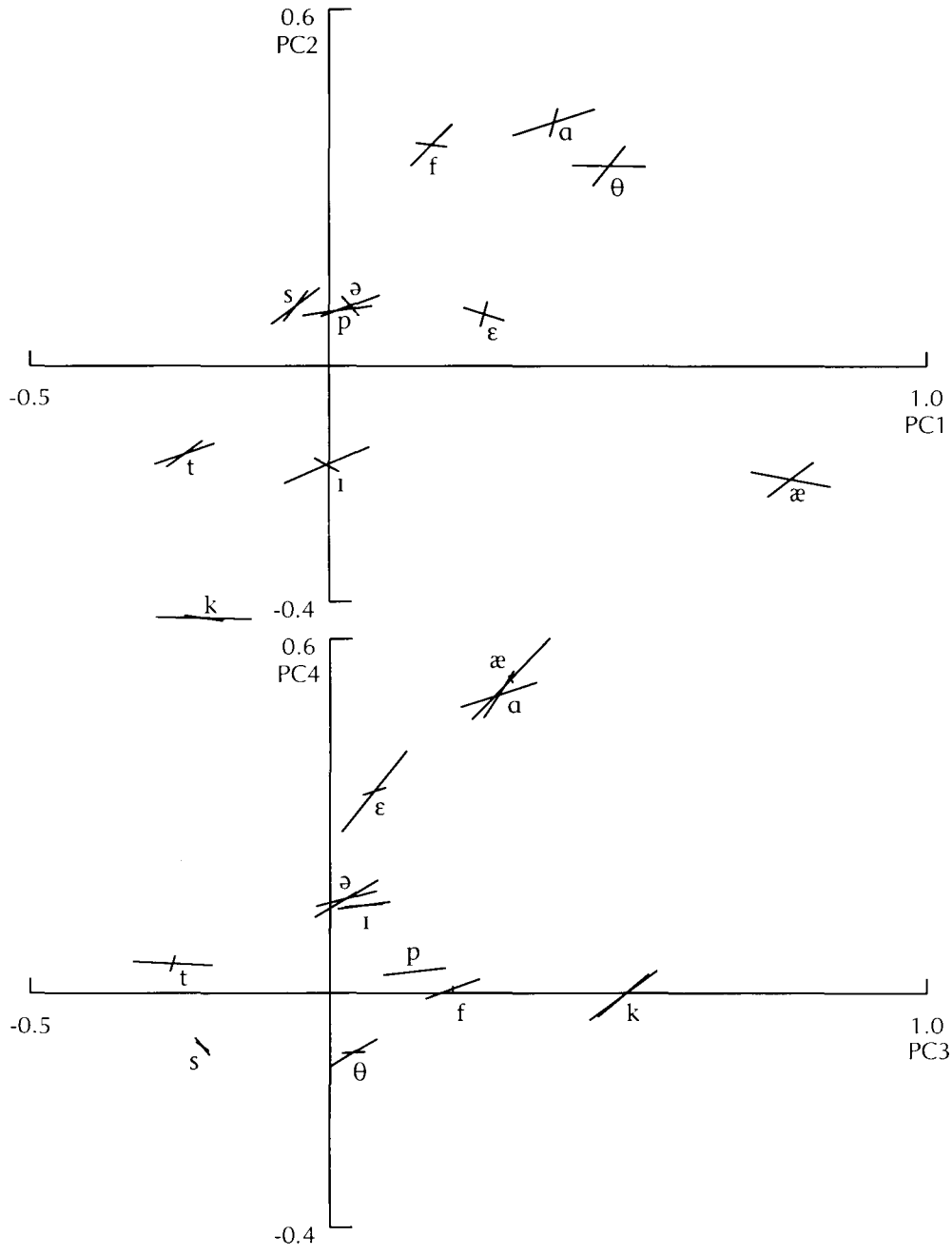


Figure A.18: Same as figure A.17—phones projected by global PCs—plotted here for experimental subject JW15. □

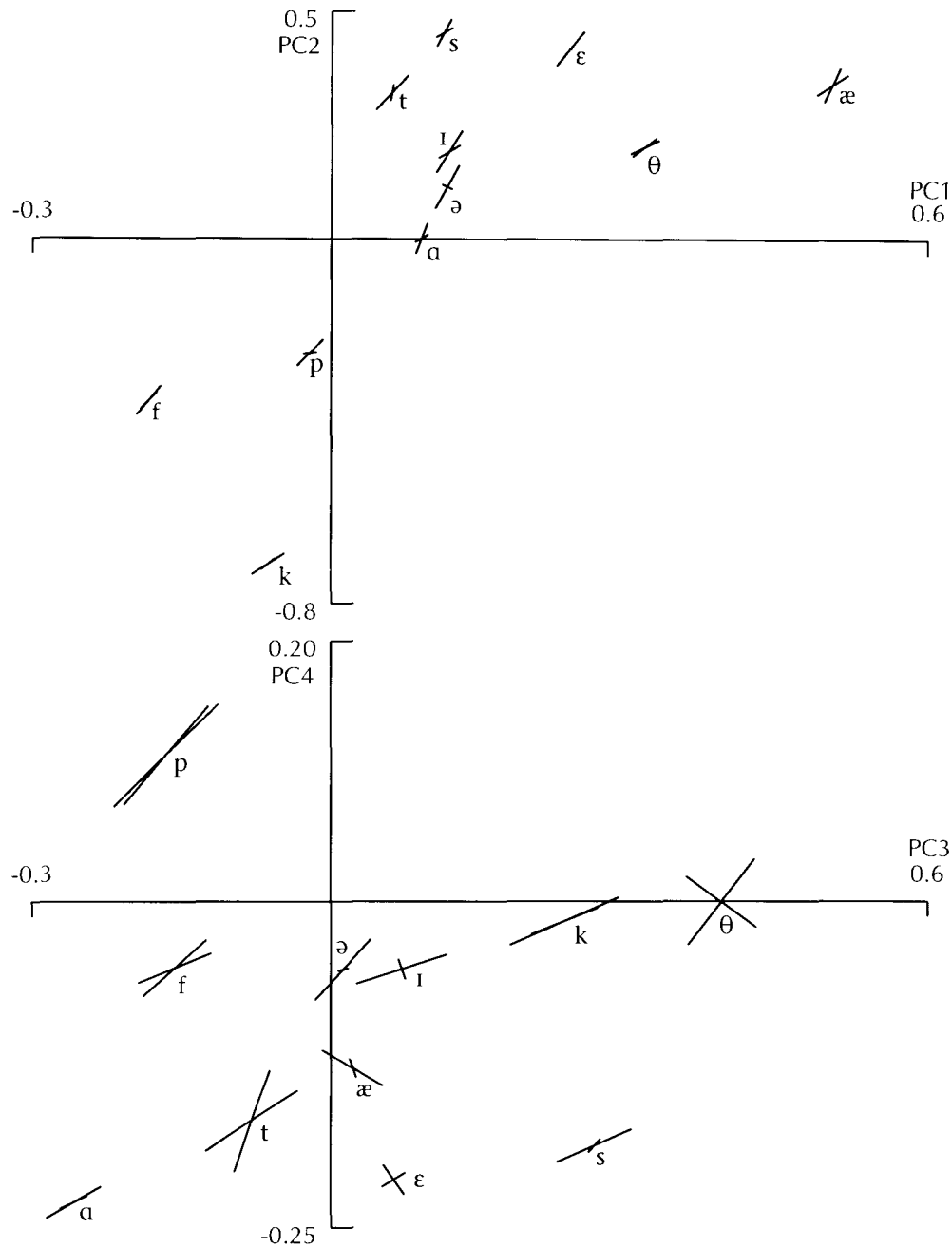


Figure A.19: Same as figure A.17 and figure A.18, plotted here for experimental subject JW27. □

From	To	Change
/f/	/s/	Lower lip moves down Tongue position changes
/ɑ/	/i/	Tongue moves up
/b/	/g/	Lips open Tongue moves up
/r/	/ð/	Tongue moves down Tongue moves forward

Table A.2: Articulatory-phonetic contrasts visible in figure A.14 through figure A.16. □

A.4 PCA: Conclusions and Discussion

In both the present thesis and the above-cited paper (Stone et al. 1997), PCA was performed globally across phonemes and a single transformation matrix was applied before classification. The primary differences (Stone et al. 1997) are:

- That project looked at sections of the tongue in a plane orthogonal to the bead coordinates used here.
- They considered only vowels, while this thesis considers 43 different phones including consonants as well.
- They analyzed data for a single speaker; recognition results are presented here for 6 speakers (chap. 3) and category statistics are plotted for 3 of those subjects (section A.3.3).
- This thesis compares joint acoustic-articulatory recognition to acoustic-only, whereas they confirmed that three broad classes of vowels could be distinguished from each other with articulation alone.
- Because this project uses longer continuous utterances instead of five-phone patterns, the process of automatically determining start and end times is more susceptible to error. This alignment/segmentation process should not be confused with recognition, which uses the same set of monophone models but lacks a priori information about the sequence in which they occur.

In both cases, the top several principal components were useful for classification, but among those few, highest variance was not synonymous with best discriminability (figure A.17). In the previous

work, the second principal component was seen to be better for classification than the first (Stone et al. 1997). In this project, qualitatively, the third and fourth together seem more useful than the first and second (figure A.17).

Appendix B

Front-End Optimization

As described in chapter 1, speech recognizers typically have a pipelined architecture consisting of a front end and a back end. Nagendra Kumar and Andreas Andreou have described a technique—heteroscedastic discriminant analysis (HDA, section B.5)—for automatically optimizing front-end processing to improve performance of the back end (Kumar 1997) (Kumar and Andreou 1998). This chapter describes a successful application of HDA to large-vocabulary, conversational speech, and reports an unforeseen problem in applying HDA to most state-of-the-art recognizers (specifically hidden Markov models using a weighted sum of Gaussians for each state's emission PDF).

B.1 Optimization of Front-End Linear Processing

Although recognizer front ends usually include both nonlinear and linear processing steps, linear optimization can be considerably easier than nonlinear. In the present project, only the linear processing steps were optimized. The space of possible transformations included linear combinations of cepstral coefficients for a given analysis frame—representing the same discretized time. It also included combinations of coefficients from different times over a wide window, special cases of which include first and second derivatives. Transformations were not constrained to be orthogonal.

B.2 Typical Front-End Components

The typical state-of-the-art set of front-end features is the Mel-frequency cepstrum with first and second time derivatives appended. This front end is described in greater detail in chapter 1. In the present context, it is worth noting that many different front ends have been tried over the years, amounting to a collective, ad hoc optimization of the front end. The techniques described here make the optimization automatic and explicit. However, since they are all restricted to linear feature transformations, they do not cover all possible front ends, or even all that have been proposed and implemented by other researchers.

B.3 Need for Dimensionality Reduction

Reducing the number of parameters produced by the front end has the following advantages:

1. Better matching of model architecture to reality, for example, by factoring out between-parameter correlations (section B.4.1).

2. Avoidance of overtraining/overfitting.

3. Reduced computational requirements.

B.4 Other Dimensionality Reduction Techniques

Neither principal components analysis (PCA, section B.4.1) nor linear discriminant analysis (LDA, section B.4.2) were used for the project of this appendix. However, consideration of these techniques may aid in understanding the heteroscedastic technique (section B.5) that was used.

B.4.1 Principal Components Analysis (PCA)

Principal components analysis (PCA) is defined in chapter 3. The fundamental difference between PCA and the discriminant techniques described below is that the latter use category information. In other words, PCA acts on a single covariance matrix, typically derived from a single set of sampled vectors $\mathbf{x}(t)$ for $1 \leq t \leq T$. The discriminant approaches add the requirement of class labels $L(t)$: $1 \leq L(t) \leq C$, identifying each sample as belonging to one of C classes.

An intermediate approach involves partitioning the data into different classes and performing PCA independently on them. Such an analysis may be used to constrain models to a subspace of the front-end parameter space. However, the dimensions perpendicular to the subspace must be treated carefully (Roweis 1999).

B.4.2 Linear Discriminant Analysis (LDA)

Linear discriminant analysis (LDA) (Fisher 1936) can be used with a labeled set of multidimensional data to find the principal directions in which the category labels differ. It is optimal, in the maximum-likelihood sense, when the variances of the different categories are identical. It has previously been applied to continuous speech recognition (Yu et al. 1990) including scaling to large-vocabulary problems (Haeb-Umbach and Ney 1992).

$$(S_B - \lambda_k S_W)\theta_k = 0$$

Equation B.1: Linear discriminant analysis (LDA) (Fisher 1936), formulated as an eigenvalue-eigenvector problem (Duda and Hart 1973). The eigenvectors θ_k are used to transform feature vectors; S_B and S_W are the between-group and within-group scatter. □

B.5 Heteroscedastic Discriminant Analysis (HDA)

In speech recognizers, multidimensional Gaussians are generally used to represent the statistics of sound classes. In particular, CDHMMs use a mixture of Gaussians to model each state's emission density (section 1.11.2). The Gaussians in the mixtures are typically constrained to have diagonal covariance matrices, but are distinct in their means and variances. A set of distributions with unequal variances are known as *heteroscedastic*. When optimizing front-end parameters for models based on such a distribution set, the technique of heteroscedastic discriminant analysis (HDA) is more appropriate than LDA.

B.5.1 Statistics Required to Perform HDA

In order to compute HDA, means and covariance matrices must be calculated for each of the categories of data, as well as for the entire data set—instead of just the two mean vectors and two covariance matrices (between-groups and within-group) which LDA requires.

The simplest way to assign classes and collect class statistics for HDA is to use a transcript with start and end times, and to assign each input frame to a particular category. Such a transcript would typically be generated via Viterbi alignment (section 1.11.2).

B.5.2 Baum-Welch Training of HDA

In order to fully integrate HDA into the HMM training procedure, the required means and covariance estimates can be borrowed from the parameter update (M) step of the EM algorithm.

B.5.3 HDA and Weighted Sums of Gaussians

Different levels of representation may be chosen for the classes which are to be differentiated by HDA. Whatever level is chosen, the class will be modeled with a Gaussian, so ideally the class should have truly Gaussian statistics. It is tempting, therefore, to assign each Gaussian in a mixture-of-Gaussians emission PDF to a separate class. Two essential problems arise:

1. HDA tries to find dimensions along which its classes differ. If mixtures are split into separate classes, instead of choosing phonetically relevant dimensions, the procedure will try to distinguish individual elements in the mixtures.

2. HDA cannot readily be extended to the mixture case, because the derivation of its objective function depends critically on statistical independence of neighboring frames and Gaussian class statistics. The former assumption allows frame probabilities to be multiplied. The latter assumption facilitates taking a logarithm, which makes frame probabilities additive, and turns the Gaussians into Mahalanobis distance calculations. Because Gaussians are combined linearly in a mixture model, taking a logarithm does not simplify the string-probability equation.

B.6 Scaling Problems in Speech Recognition

New approaches to speech recognition that show promise on small data sets often cannot be applied successfully to larger problems. The present project tested whether the HDA technique actually scaled to continuous, large-vocabulary speech recognition.

B.7 Switchboard Telephone Corpus

The Switchboard data set (Godfrey et al. 1992), used for this project, includes roughly 250 hours of voluntarily recorded telephone conversations. The conversations took place between pairs of experimental subjects and were initiated by computer prompts suggesting a topic. A word-level (as opposed to phone-level) transcription is included with the data. The corpus has been modified over the years to correct errors in the text.

B.7.1 Large-Vocabulary Continuous Speech Recognition Workshop

The results reported in this chapter were obtained at the annual large-vocabulary speech recognition workshop at the Johns Hopkins University, which is attended by an international group of speech-recognition researchers in academia, industry, and government. The workshop is a competition for the greatest reduction in error rate from an agreed-upon baseline (Jelinek 1996). The results reported here were the second best for that year's workshop (Andreou et al. 1998).

B.8 Experimental Procedure

Application of HDA to transcription of conversations involved the following steps:

1. Generate overlapping context windows, each including feature vectors for nine successive cepstral-analysis frames;
2. Collect monophone statistics from triphone alignment and overlapping-windowed data;
3. Use HDA, implemented via numerical optimization, to find a dimensionality-reducing transformation from monophone statistics;
4. Use single-pass retraining to convert models to new feature set;
5. Perform several additional E-M iterations;
6. Test models.

Training used approximately 114,000 utterances from the revised Switchboard data set. Each utterance represented one speaker's turn in conversation, so length varied from single-word interjections to multiple-sentence comments.

B.8.1 Generation of Context Windows

Along with each cepstral vector, the four preceding and the four following vectors were concatenated. This gave a sequence of overlapping context windows, each having nine times as many coefficients as the original front end.

B.8.2 Segmentation per Triphone Alignment

Transcripts with start and end times for each triphone were provided by Dr. Bill Byrne. For HDA, triphone contexts were pooled so that the classes would represent monophones. Triphone HMMs remained distinct.

The time-aligned monophone transcripts were used to collect all extended vectors corresponding to each monophone. From these sets, a mean and a covariance matrix were generated for each monophone.

B.8.3 Running HDA

The statistics were input to a numerical optimization implementation of HDA, which gave a transformation projecting the context windows into a 39-dimensional space. Due to software limitations, the latter number of dimensions had to match the number of features in the conventional front end (12 Mel-frequency cepstral coefficients plus energy, and their deltas and accelerations). For comparison, a dimensionality-reducing transformation was also obtained using LDA.

B.8.4 Retraining

Models which had been partially trained using the conventional front end were converted to the new feature set using HTK's single-pass retraining (Young et al. 1997). In this procedure, the forward-backward (expectation) step is performed in the old parameter space, but the parameter estimation (maximization) step is performed for the new features. The forward-backward algorithm gives the probability $P_B(q,t)$ that a state q will be occupied at each time t; these probabilities scale the contribution of reparameterized input vector t to the emission statistics of state q. After single-pass retraining, several additional E-M steps are performed in the new space.

Processing	Word error rate
Baseline	49.9%
LDA	51.1
HDA	49.1

Table B.1: Word error rates on Switchboard data, for baseline recognizer, LDA transformation, and HDA transformation (Fain et al. 1997) □

B.8.5 Problems with Variance Floor and Grammar Weight

The new parameterization may have resulted in a new typical magnitude for the feature vectors. Since emission probability densities have units inversely proportional to feature-vector units, the probability densities would scale inversely. This, in turn, suggests that the variance floor (section 3.11.1) and grammar weight (section 3.11.2) might no longer have been appropriate. Both parameters had been carefully tuned for the old feature set. So while the baseline recognition results represented optimal values, the HDA results may have been suboptimal—and the following estimate of the advantage of using HDA may be conservative.

B.8.6 Results

Results of applying LDA and HDA to transcription of telephone conversations appear in table B.1. Evaluation used a held-out test set, so no utterances used in training were used for testing. While LDA caused an increase in errors, HDA reduced the number of errors by 1.6% relative to the baseline (Fain et al. 1997). This improvement is typical of the most successful techniques at each year's workshop (section B.7.1).

B.9 Conclusions

Heteroscedastic discriminant analysis improves performance on telephone-conversation transcription, compared to a conventional architecture. The observed improvement, a 1.6% reduction in error rate (section B.8.6) (Fain et al. 1997), is probably a conservative estimate due to suboptimal choice of some global parameters (section B.8.5). Further improvement is possible by integrating HDA fully into the E-M training procedure (section B.5.2).

Appendix C

Lipreading Aids for the Hearing Impaired

C.1 Lipreading by the Hearing Impaired

It is widely known that the hearing impaired use lipreading to help understand speech, so facial animation or other visual cues, derived from sound, may compensate for some hearing loss.

Hubert Upton built modified eyeglasses to aid the hearing impaired in lipreading (Upton 1968). The glasses included light-emitting diodes (LEDs), a microphone, and analog circuitry to determine what general class of speech sound was being picked up by the microphone. This category information (voiced/unvoiced and fricative/stop/other) was displayed by LEDs in positions related to speech production. For example, an LED placed near the bottom of the glasses (to represent the role of the vocal cords) indicated the voiced/unvoiced distinction.

Another type of lipreading aid performs phone-level recognition from the telephone line and animates a lip image (Slager 1993) (Carraro et al. 1989). A newer facial animation system for the hearing impaired was found to reduce recognition errors in hearing-impaired subjects (Agelfors et al. 1998). With natural speech, the synthetic animated face reduced word errors by 21% compared to sound alone. When video of the speaker was used, the reduction was 74%. In a restricted

context—recognition of isolated vowel-consonant-vowel sequences—the error rate reductions were 36% and 40%, respectively.

Appendix D

Parameterizations of Lip and Jaw Motion in FACS and MPEG-4

Lipreading (despite the name) can also involve perception of tongue motion. In the side view project of chapter 2, the tongue is fairly hard to see, so the effective degrees of freedom for a particular speaker and context can be estimated by considering the lips and jaw.

D.1 Facial Action Coding System (FACS)

The facial action coding system (FACS) (Ekman and Friesen 1978) classifies motions of the face, including those used for speech and the expression of the emotions. Each action may involve multiple muscles, and a facial expression may involve multiple actions. Those involving the lips and jaw appear in table D.1. This set gives a very rough estimate of the number of degrees of freedom of motion of those articulators—19.

D.2 Motion Picture Experts Group (MPEG) 4, Synthetic-Natural Hybrid Coding (SNHC)

A newer standard for facial expression coding is part of the Motion Picture Experts' Group (MPEG) 4 standard for Synthetic-Natural Hybrid Coding (SNHC) (MPEG-4 SNHC 1996). The goal of SNHC is to mix video and audio recordings with animation and synthesis. Initially, this might be as simple as overlaying computer-animated characters and a soundtrack of electronic music on a film of live actors. The synthetic component is very flexible, providing a programming language for representing and transmitting the decoding technique.

The face and body definition parameters (FDP) and facial animation parameters (FAP) drive character animation, with FDP describing the geometry of a face, and FAP used to describe motion.

Articulator(s)	**Action**	**Muscles**
Both lips:	toward each other	Orbicularis oris
	pull corner (left/right)	Zygomatic major
	depress corner (left/right)	Triangularis
	suck	Orbicularis oris
	pucker	Incisivii labii superioris Incisivii labii inferioris
	stretch	Risorius
	funnel	Orbicularis oris
	tighten	Orbicularis oris
	press	Orbicularis oris
	part	Depressor labii *or* Relaxation of mentalis *or* Relaxation of orbicularis oris
	blow/puff	
Upper lip:	raise	Levator labii superioris Caput infraorbitalis
Lower lip:	depress	Depressor labii
	bite	
Jaw/chin:	raise	Mentalis
	drop	Masseter Temporal and internal pterygoid
	thrust	
	move sideways	
	clench	

Table D.1: Lip and jaw motions in the facial action coding system (Ekman and Friesen 1978). □

Appendix E

X-Ray Microbeam Tracking Technology

The articulator-motion data used in the experiments of chapter 3 were obtained, in previous work, using an x-ray microbeam tracking system at the University of Wisconsin (Westbury et al. 1994). This system enables occluded movements (e.g., inside the mouth) to be observed with high time resolution—recording an updated position about every 10 ms. The Wisconsin system was based upon an earlier apparatus developed at the University of Tokyo (Fujimura et al. 1973).

E.1 Microbeam Tracking as a Substitute for Cineradiography

Microbeam recording was developed in response to cineradiography (chap. 3), in which motion pictures of subjects talking were filmed using x-rays. Cineradiography caused serious health concerns, since the subject's head was exposed on the order of every 15 ms. In microbeam tracking, radiation is concentrated in a small beam surrounding the target. This reduces radiation exposure by about three orders of magnitude (Fujimura et al. 1973). In the Wisconsin experiments, the targets were gold beads of 2-3 mm diameter, and the irradiated beam had a cross section of about 6 mm square.

E.2 Tracking Algorithm

The small irradiated areas follow the beads as they move. A computer determines the beads' positions within the irradiated areas and predicts their positions during the next time step. At the next sampling time, a search for each bead is performed in the predicted area. If a bead is not found near its predicted location, the search expands to cover a wider range of locations.

E.3 Bead Placement

The Wisconsin system tracks a number of beads simultaneously by alternating between them. Eight beads are placed in positions relevant to articulation—four on the tongue, one on each lip, and two on the lower jaw—and two (Westbury 1991) or three (Westbury et al. 1994) beads are used as a reference for head motion.

Appendix F

Text Samples

F.1 Paragraphs used for Recognition Experiments

The University of Wisconsin x-ray microbeam data set contains the following paragraphs (Westbury et al. 1994); they were used for the paragraph-recognition experiments of chapter 3. The task numbers given below are with reference to the entire data set, which also includes word and sentence tasks.

F.1.1 "Grandfather" Paragraphs

The W. B. Saunders Company holds the copyright on the first two paragraphs used for recognition (Darley et al. 1975). They requested that the text not be distributed electronically. In light of that request, the fair use doctrine will not be invoked and the paragraphs have been omitted. The paragraphs appear as task 11 and task 12 in the database description (Westbury et al. 1994).

F.1.2 "Hunter" Passage

Reprinted with permission from T. H. Crystal and A. H. House (1982). "Segmental durations in connected speech signals: preliminary results." *Journal of the Acoustical Society of America* 72. p. 715. Copyright 1982, Acoustical Society of America.

Paragraph 1 (Task 78)

> In late fall and early spring the short rays of the sun call a true son of the out-of-doors back to the places of his childhood. Tom Brooks was such a man. Each year at these times his desk seemed like a stone whose weight made him wish for the life he knew as a boy. In the five years since leaving college he had not revisited his old haunts before. But this March Tom found himself by a small stream with a gun.

Paragraph 2 (Task 79)

> This March, Tom found himself by a small stream with a gun at rest in the crook of his arm. The desk that had tied him down was gone and his one thought was for quail. He had been on the trail since dawn, but not one bird had crossed his path. It seemed as though five years without hunting had made him lose touch with all the small signs that he once knew—signs that would tell for sure if an animal was near or not.

Paragraph 3 (Task 80)

> Once he thought he saw a bird, but it was just a large leaf that had failed to drop to the ground during the winter. Tom stopped near a small stream to rest. Soon after he had laid down his gun, he heard the sound of wings from across the stream, and five large birds came out of the brush. They flew to the edge of the stream unaware of the hunter.

Paragraph 4 (Task 81)

> The birds flew to the edge of the stream unaware of the hunter. Tom placed his hand on his gun quietly. Slowly he raised it to his shoulder and took aim. The seconds ticked off like hours, but still the birds drank. Quick shots rang out. The years of waiting seemed to disappear with the successful culmination of the hunt.

F.2 Example Sentences for Importance of Consonants and Vowels

The sentences quoted in chapter 4 with letters removed were taken a textbook on Old English, where they appeared as examples of text grammatical in both Old and Modern English (Mitchell and Robinson 1991). In full, they read:

Grind his corn for him and sing me his song.

He swam west in storm and wind and frost.

F.3 Random Text Generated by a Markov Model

A combined high-order and first-order Markov chain can be used as a text generator (Raymond 1993) (Stallman 2000). The high-order model repeats sections of training text, and periodically the first-order model jumps to a new section based on word-to-word transition probability estimates. Trained on a draft of this thesis, such a generator results in the following nonsensical yet vaguely plausible random text:

> Coarticulation with vowels can be addressed by normalizing each component of the feature vector to the corresponding word in the true text of percent correct to evaluate recognition performance. This choice is motivated by the neck axis.

References

Adjoudani, A. and Benoît, C. 1996. "On the integration of auditory and visual parameters in an HMM-based ASR," in *Speechreading by Humans and Machines: Models, Systems, and Applications*, Stork, D. G. and Hennecke, M. E., editors, pages 461–471. Springer-Verlag.

Agelfors, E., Beskow, J., Dahlquist, M., Granström, B., Lundeberg, M., Spens, K.-E., and Öhman, T. 1998. "Synthetic faces as a lipreading support," in *Proceedings of the Fifth International Conference on Spoken Language Processing (ICSLP)*, Sydney, Australia. Australian Speech Science and Technology Association.

Ames, C., 1989. "The Markov process as a compositional tool: A survey and tutorial," *Leonardo*, 22(2):175–188.

Andreou, A. G., Hermansky, H., Wellekens, C. J., Minami, Y., Kamm, T., Luettin, J., Fain, D. C., and van Vuuren, S. 1998. "Acoustic processing group, WS97: Final report," in *1997 Large Vocabulary Continuous Speech Recognition Summer Research Workshop Technical Reports*, Jelinek, F., editor. Johns Hopkins University Center for Language and Speech Processing, Baltimore, MD, USA.

Atal, B. S., Chang, J. J., Mathews, M. V., and Tukey, J. W., 1978. "Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer-sorting technique," *Journal of the Acoustical Society of America*, 63(5):1535–1555.

Baer, T., Gore, J. C., Gracco, L. C., and Nye, P. W., 1991. "Analysis of vocal-tract shape and dimensions using magnetic resonance imaging: Vowels," *Journal of the Acoustical Society of America*, 90(2):799–828. Part One.

Bahl, L. R., Brown, P. F., deSouza, P. V., and Mercer, L. R. 1986. "Maximum mutual information estimation of hidden Markov model parameters for speech recognition," in *IEEE International*

Conference on Acoustics, Speech, and Signal Processing (ICASSP), pages 49–52, Tokyo, Japan. IEEE.

Bailey, P. J. 1983. "Hearing for speech: The information transmitted in normal and impaired hearing," in *Hearing and Hearing Disorders*, Lutman, M. E. and Haggard, M. P., editors. Academic Press, London, England, UK.

Bakis, R. 1974. "Continuous-speech word spotting via centisecond acoustic states," Technical Report RC 4788, International Business Machines, Yorktown Heights, NY, USA.

Bass, L., Mann, S., Siewiorek, D., and Thompson, C., 1997. "Issues in wearable computing: A CHI 97 workshop," *SIGCHI Bulletin*, 29(4).

Baum, L. E., 1972. "An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes," *Inequalities*, 3:1–8.

Bell, A. M. 1867. *The Visible Speech: The Science of Universal Alphabetics.*

Bengio, Y. and de Mori, R. 1988. "Use of neural networks for the recognition of place of articulation," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 103–106. IEEE.

Benoît, C., Guiard-Marigny, T., Goff, B. L., and Adjoudani, A. 1996. "Which components of the face do humans and machines best speechread?," in *Speechreading by Humans and Machines: Models, Systems, and Applications*, Stork, D. G. and Hennecke, M. E., editors, pages 315–328. Springer-Verlag.

Blackburn, S. 1997. *Articulatory Methods for Speech Production and Recognition.* PhD thesis, Cambridge University, Trinity College, Cambridge, England, UK.

Bogert, B. P., Healy, M. J. R., and Tukey, J. W. 1963. "The quefrency alanysis [sic] of time series for echoes: Cepstrum, pseudo-autocovariance, cross-cepstrum, and saphe [sic] cracking," in *Proceedings of the Symposium on Time Series Analysis*, Rosenblatt, M., editor, pages 209–243, New York, NY, USA. John Wiley and Sons.

Bourlard, H., Hermansky, H., and Morgan, N., 1996. "Towards increasing speech recognition error rates," *Speech Communication*, 18:205–231.

Bourlard, H., Konig, Y., and Morgan, N. 1994. "REMAP: recursive estimation and maximization of a posteriori probabilities—application to transition-based connectionist speech recognition," Technical Report TR-94-064, International Computer Science Institute, Berkeley, CA, USA.

Breen, A., 1992. "Speech synthesis models: a review," *Electronics and Communication Engineering Journal*, pages 19–31.

Broca, P., 1865. "Sur le siége de la faculté du langage articulé," *Bulletins de la Société d'anthropologie de Paris*, 6:377–393.

Brooks, F., Hopkins, A., Newmann, P., and Wright, W., 1957. "An experiment in musical composition," *Institute of Radio Engineers Transactions on Electronic Computers*, EC-6(1):175–182.

Browman, C. P. and Goldstein, L., 1989. "Articulatory gestures as phonological units," *Phonology*, 6:201–251.

Browman, C. P. and Goldstein, L., 1992. "Articulatory phonology: an overview," *Phonetica*, 49:155–180.

Brown, P. F., Della Pietra, S. A., Della Pietra, V. J., Lai, J. C., and Mercer, R. L., 1992. "An estimate of an upper bound for the entropy of English," *Computational Linguistics*, 18(1):31–40.

Byrne, W., Finke, M., Khudanpur, S., McDonough, J., Nock, H., Riley, M., Saraclar, M., Wooters, C., and Zavaliagkos, G. 1998. "Pronunciation modeling at WS97," in *1997 Large Vocabulary Continuous Speech Recognition Summer Research Workshop Technical Reports*, Jelinek, F., editor. Johns Hopkins University Center for Language and Speech Processing, Baltimore, MD, USA.

Calvert, G. A., Bullmore, E. T., Brammer, M. J., Campbell, R., Williams, S. C. R., McGuire, P. K., Woodruff, P. W. R., Iverson, S. D., and David, A. S., 1997. "Activation of auditory cortex during silent lipreading," *Science*, 276:593–596.

Carraro, A., Chilton, E. H. S., and McGurk, H. 1989. "A telephonic lipreading device for the hearing impaired," in *IEE Colloquium on Biomedical Applications of Digital Signal Processing*, Stevenage, UK. IEE.

Cathiard, M. A., Cirot-Tseva, A., and Lallouache, M. T., 1992. "Identification visuelle des gestes de protrusion et de rétraction des lèvres au cours des pauses acoustiques: Les performances de sujets français et grecs," *Journal de Physique IV*, 2:C1–319–C1–322. Colloque C1 supplément.

Chang, J. W. and Glass, J. R. 1997. "Segmentation and modeling in segment-based recognition," in *Eurospeech*, volume 3, pages 1199–1202.

Chomsky, N. and Halle, M. 1968. *The Sound Pattern of English*. Harper and Row, New York, NY, USA.

Cohen, D., 1972. "Magnetoencephalography: Detection of the brain's electrical activity with a superconducting magnetometer," *Science*, 175:664–666.

Cohen, L. 1995. *Time-frequency Analysis*. Prentice Hall PTR, Englewood Cliffs, NJ, USA.

Cole, R. A., Yan, Y., Mak, B., Fanty, M., and Bailey, T. 1996. "The contribution of consonants versus vowels to word recognition in fluent speech," in *International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 853–856. IEEE.

Coleman, J., 1998. "Cognitive reality and the phonological lexicon: A review," *Journal of Neurolinguistics*, 11(3):295–320.

Cover, T. M. and King, R. C., 1978. "A convergent gambling estimate of the entropy of English," *IEEE Transactions on Information Theory*, IT-24(4):413–421.

Crystal, T. H. and House, A. S., 1982. "Segmental durations in connected speech signals: Preliminary results," *Journal of the Acoustical Society of America*, 72:705–716.

Darley, F. L., Aronson, A. E., and Brown, J. R. 1975. *Motor Speech Disorders*. W. B. Saunders, Philadelphia, PA, USA.

Davis, H. K., Biddulph, R., and Balashek, S., 1952. "Automatic recognition of spoken digits," *Journal of the Acoustical Society of America*, 24:637–642.

Davis, S. B. and Mermelstein, P., 1980. "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(4):357–366.

Deken, J., 1979. "Some limit results for longest common subsequences," *Discrete Mathematics*, 26:17–31.

Dekle, D. J., Fowler, C. A., and Funnell, M. G., 1992. "Audiovisual integration in perception of real words," *Perception and Psychophysics*, 51(4):355–362.

Delattre, P. C., Liberman, A. M., and Cooper, F. S., 1955. "Acoustic loci and transitional cues for consonants," *Journal of the Acoustical Society of America*, 27:769–773.

Demolin, D., George, M., Lecuit, V., Metens, T., Soquet, A., and Raeymaekers, H. 1997. "Coarticulation and articulatory compensations studied by dynamic MRI," in *Eurospeech*, volume 1, pages 43–46.

Dempster, A. P., Laird, N. M., and Rubin, D. B., 1977. "Maximum likelihood estimation from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society, Series B*, 39:1–38.

Denes, P. and Matthews, M. V., 1960. "Spoken digit recognition using time-frequency pattern matching," *Journal of the Acoustical Society of America*, 32:1450–1455.

Denes, P. B., 1963. "On the statistics of spoken English," *Journal of the Acoustical Society of America*, 35(6):892–904.

Deng, L., Kenny, P., Lennig, M., and Mermelstein, P., 1992. "Modeling acoustic transitions in speech by state-interpolation hidden Markov models," *IEEE Transactions on Signal Processing*, 40(2).

Deng, L. and Sun, D. X., 1994. "A statistical approach to automatic speech recognition using the atomic speech units constructed from overlapping articulatory features," *Journal of the Acoustical Society of America*, 95(5):2702–2719. Part One.

Duda, R. O. and Hart, P. E. 1973. *Pattern Classification and Scene Analysis*. John Wiley and Sons.

Dudley, H., 1940. "The carrier nature of speech," *Bell Systems Technical Journal*, 19(4):495–515.

Dudley, H., Riesz, R. R., and Watkins, S. A., 1939. "A synthetic speaker," *Journal of the Franklin Institute*, pages 739–764.

Easton, R. D. and Basala, M., 1982. "Perceptual dominance during lipreading," *Perception and Psychophysics*, 32:562–570.

Eimas, P. D., Miller, J. L., and Jusczyk, P. W. 1987. "On infant speech perception and the acquisition of language," in *Categorical Perception: The Groundwork of Cognition*, Harnad, S., editor. Cambridge University Press, New York, NY, USA.

Ekman, P. and Friesen, W. 1978. *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Consulting Psychologists Press, Palo Alto, CA, USA.

Erler, K. and Freeman, G. H., 1996. "An HMM-based speech recognizer using overlapping articulatory features," *Journal of the Acoustical Society of America*, 100(4):2500–2513. Part One.

Fain, D. C. 1997. "Imaging system and lipreading technique for cockpit voice recognition: Proposal,"

Fain, D. C. 1998. "Imaging system and lip-reading technique for cockpit voice recognition: Final report,"

Fain, D. C., Andreou, A. G., and Kamm, T. 1997. "Heteroscedastic discriminant analysis: Optimal feature filtering," in *Proceedings of the 1997 Summer Workshop on Large Vocabulary Continuous Speech Recognition*, Baltimore, MD, USA. Johns Hopkins University Center for Language and Speech Processing.

Fain, D. C. and Chinn, G. 1999. "Imaging system and lip-reading technique," U. S. Patent Filing—Docket 33956/SAH/T326. 1999.

Fant, G. 1970. *Acoustic Theory of Speech Production*. Mouton, The Hague, Netherlands.

Ferretti, M., Maltese, G., and Scarci, S., 1990. "Measuring information provided by language model and acoustic model in probabilistic speech recognition: Theory and experimental results," *Speech Communication*, 9:531–539.

Fisher, R. A., 1936. "The use of multiple measurements in taxonomic problems," *Annals of Eugenics*, 7:179–188.

Fonollosa, J. R., 1996. "Positive time-frequency distributions based on joint marginal constraints," *IEEE Transactions on Signal Processing*, 44(8):2086–2091.

Fry, D. B. and Denes, P., 1958. "The solution of some fundamental problems in mechanical speech recognition," *Language and Speech*, 1:35–58.

Fujimura, O., Kiritani, S., and Ishida, H., 1973. "Computer-controlled radiography for observation of movements of articulatory and other human organs," *Computers in Biology and Medicine*, 3:371–384.

Furui, S., 1986. "Speaker independent isolated word recognition using dynamic features of speech spectrum," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 34(1):52–59.

Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S., and Dahlgren, N. L. 1993. "DARPA TIMIT acoustic-phonetic continuous speech corpus," Technical Report NISTIR 4930, Computer Systems Laboratory, National Institute of Standards and Technology (NIST), Gaithersburg, MD, USA. http://www.ldc.upenn.edu.

Giguere, C., Bosman, A. J., and Smoorenburg, G. F., 1997. "Automatic speech recognition experiments with a model of normal and impaired peripheral hearing," *Acustica*, 83(6):1065–1076.

Godfrey, J., Holliman, E., and McDaniel, J. 1992. "Switchboard: Telephone speech corpus for research and development," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 517–520, San Francisco, CA, USA. http://www.ldc.upenn.edu.

Greenwood, A. R., Goodyear, C. C., and Martin, P. A., 1992. "Measurements of vocal-tract shape using magnetic resonance imaging," *IEE Proceedings I: Communications, Speech, and Vision*, 139(6):553–560.

Haeb-Umbach, R. and Ney, H. 1992. "Linear discriminant analysis for improved large vocabulary continuous speech recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pages 13–16, San Francisco, CA, USA.

Hahn, K. 1994. "Frequency of English letters (crypto-stats) from *A Tale of Two Cities*," Posted from Loral Data Systems to `sci.crypt` newsgroup. 1994.

Hari, R., 1991. "Activation of the human auditory cortex by speech sounds," *Acta Oto-Laryngologica*, Suppl. 491:132–138.

Harnad, S. 1987. "Psychophysical and cognitive aspects of categorical perception: A critical overview," in *Categorical Perception: The Groundwork of Cognition*, Harnad, S., editor. Cambridge University Press, New York, NY, USA. Includes discussion of motor theory of human speech perception.

Harris, C. M., 1953. "A study of the building blocks in speech," *Journal of the Acoustical Society of America*, 25:962–969.

Harshman, R., Ladefoged, P., and Goldstein, L., 1977. "Factor analysis of tongue shapes," *Journal of the Acoustical Society of America*, 62(3):693–707.

Hennecke, M. E., Stork, D. G., and Prasad, K. V. 1996. "Visionary speech: Looking ahead to practical speechreading systems," in *Speechreading by Humans and Machines: Models, Systems, and Applications*, Stork, D. G. and Hennecke, M. E., editors, pages 331–349. Springer-Verlag.

Hogden, J., Lofquist, A., Gracco, V., and Oshima, K. 1993. "Inferring articulator positions from acoustics: an electromagnetic midsagittal articulometer experiment," in *One Hundred Twenty-sixth Meeting of the Acoustical Society of America*, volume 94, page 1764.

Holmgren, K., Lindblom, B., Aurelius, G., Jalling, B., and Zetterstrom, R. 1986. "On the phonetics of infant vocalization," in *Precursors of Early Speech*, Lindblom, B. and Zetterstrom, R., editors, pages 51–66. Academic Press.

Höwing, F., Wermser, D., and Dooley, L. S., 1996. "Recognition and tracking of articulatory organs in x-ray image sequences," *Electronics Letters*, 32(5):444–445.

Imaizumi, S., Mori, K., Kiritani, S., Hosoi, H., and Tonoike, M., 1998. "Task-dependent laterality for cue decoding during spoken language," *Neuroreport*, 9:899–903.

Jakobson, R., Fant, G., and Halle, M. 1952. "Preliminaries to speech analysis: The distinctive features and their correlates," Technical Report 13, Acoustics Laboratory, Massachusettes Institute of Technology, Cambridge, MA, USA.

Jelinek, F., 1996. "Five speculations (and a divertimento) on the themes of H. Bourlard, H. Hermansky, and N. Morgan," *Speech Communication*, 18(3):242–246.

Jelinek, F., editor. 1998. *1997 Large Vocabulary Continuous Speech Recognition Summer Research Workshop Technical Reports*. Johns Hopkins University Center for Language and Speech Processing, Baltimore, MD, USA.

Jelinek, F. 1998. *Statistical Methods for Speech Recognition*. MIT Press, Cambridge, MA, USA.

Jespersen, O. 1914. *Lehrbuch der Phonetik*. Teubner, Leipzig.

Jordan, M. I. and Rosenbaum, D. 1989. "Action," in *Foundations of Cognitive Science*, Posner, M. I., editor. MIT Press, Cambridge, MA, USA.

Kamm, T., Andreou, A. G., and Hermansky, H. 1997. "Learning the Mel scale and optimal VTN [vocal tract normalization] mapping," in *Proceedings of the 1997 Summer Workshop on Large Vocabulary Continuous Speech Recognition*, Baltimore, MD, USA. Johns Hopkins University Center for Language and Speech Processing.

Kandel, E. R., Schwartz, J. H., and Jessell, T. M. 1991. *Principles of Neural Science*. Elsevier Science, New York, NY, USA.

Kass, M., Witkin, A., and Terzopoulos, D. 1987. "Snakes: Active contour models," in *Proceedings of the First International Conference on Computer Vision*.

Kent, R. D. and Minifie, F. D., 1977. "Coarticulation in recent speech production models," *Journal of Phonetics*, 5:115–135.

Kim-Renaud, Y.-K. 1997. "A brief description of the Korean alphabet," in *The Korean Alphabet: its History and Structure*, Kim-Renaud, Y.-K., editor, pages 279–287. University of Hawaii Press, Honolulu, HI, USA.

Knudsen, E. I., 1982. "Auditory and visual maps of space in the optic tectum of the owl," *Journal of Neuroscience*, 2:1117–1194.

Kuhl, P. K., Williams, K. A., Lacerda, F., Stevens, K. N., and Lindblom, B., 1992. "Linguistic experience alters phonetic perception in infants by six months of age," *Science*, 255:606–608.

Kumar, N. 1997. *Investigation of Silicon Auditory Models and Generalization of Linear Discriminant Analysis for Improved Speech Recognition*. PhD thesis, Johns Hopkins University, Baltimore, MD, USA.

Kumar, N. and Andreou, A. G., 1998. "Heteroscedastic discriminant analysis and reduced rank HMMs for improved speech recognition," *Speech Communication*, 26(4):283–297.

Labov, W. 1996. "The organization of dialect diversity in North America," in *Fourth International Conference on Spoken Language Processing*, Philadelphia. http://www.ling.upenn.edu/phono_atlas/.

Lakshminarayanan, A. V., Lee, S., and McCutcheon, M. J., 1991. "MR imaging of the vocal tract during vowel production," *Journal of Magnetic Resonance Imaging*, 1:71–76.

Larar, J. N., Schroeter, J., and Sondhi, M. M., 1988. "Vector-quantization of the articulatory space," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-36(12):1812–1818.

Lee, K.-F. and Hon, H.-W. 1988. "Speaker-independent phone recognition using hidden Markov models," Technical report, Carnegie-Mellon University, Pittsburgh, PA, USA.

Lee, K.-F., Hon, H.-W., and Reddy, R., 1990. "An overview of the Sphinx speech recognition system," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 38(1):35–45.

Levenshtein, V. I., 1966. "Binary codes capable of correcting deletions, insertions, and reversals," *Cybernetics and Control Theory*, 10(8):707–710. Original Russian-language publication in 1965.

Liberman, A. and Mattingly, I., 1985. "The motor theory of speech perception revised," *Cognition*, 21:1–36.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., and Studdert-Kennedy, M., 1967. "Perception of the speech code," *Psychological Review*, 74:431–461.

Liberman, A. M. and Whalen, D. H., 2000. "On the relation of speech to language," *Trends in Cognitive Sciences*, 4:187–196.

Liljenkrants, J. 1971. "A Fourier series description of the tongue profile," Technical Report QPSR 4, Speech Transmission Lab, Stockholm, Sweden.

Locke, J. L., 1992. "Structure and stimulation in the ontogeny of spoken language," *Developmental Psychobiology*, 28:430–440.

Loughlin, P. J., Pitton, J. W., and Atlas, L. E., 1994. "Construction of positive time-frequency distributions," *IEEE Transactions on Signal Processing*, 42(10):2697–2705.

MacDonald, J. and McGurk, H., 1978. "Visual influences on speech perception processes," *Perception and Psychophysics*, 24(3):253–257.

MacKay, D. M., 1951. "Mindlike behaviour in artefacts," *British Journal for the Philosophy of Science*, 2:105–121. Reception and replication approaches to machine perception defined; articulatory recognition an example of the latter.

Massaro, D. W. 1996. "Bimodal speech perception: A progress report," in *Speechreading by Humans and Machines: Models, Systems, and Applications*, Stork, D. G. and Hennecke, M. E., editors, pages 79–101. Springer-Verlag.

McGurk, H. and MacDonald, J., 1976. "Hearing lips and seeing voices," *Nature*, 264:746–748.

Merriam-Webster. 1996. *Collegiate Dictionary*. Springfield, MA, USA, tenth edition.

Mitchell, B. and Robinson, F. C. 1991. *A Guide to Old English*. Blackwell.

MPEG-4 SNHC. 1996. "Motion Picture Experts Group 4 Synthetic-Natural Hybrid Coding: Face and body definition and animation parameters," Technical Report JTC1/SC29/WG11 MPEG96/N1365, International Standards Organization (ISO) and International Electrotechnical Commission (IEC).

Munhall, K. G., Vatikiotis-Bateson, E., and Tohkura, Y. 1994. "X-ray film database for speech research," Technical Report TR-H-116, ATR Human Information Processing Research Laboratories, Kyoto, Japan.

Myers, C. S. and Rabiner, L. R., 1981. "A level building dynamic time warping algorithm for connected word recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-29(2):284–297.

Narayanan, S. S., Alwan, A. A., and Haker, K., 1995. "An articulatory study of fricative consonants using magnetic resonance imaging," *Journal of the Acoustical Society of America*, 98(3):1325–1347.

Nassimbene, E. G. 1965. "Electronic lip reader," U. S. Patent 3,192,321. 1965.

Nocerino, N., Soong, F. K., Rabiner, L. R., and Klatt, D. H., 1985. "Comparative study of several distortion measures for speech recognition," *Speech Communication*, 4:317–331.

Orr, D. B., Friedman, H. L., and Williams, J. C. C., 1965. "Trainability of listening comprehension of speeded discourse," *Journal of Educational Psychology*, 56:148–156.

OUP. 1989. *Oxford English Dictionary*. Oxford University Press, Oxford, England, UK, second edition.

Paget, R. A. S. 1922. "Improvements in the method of and in apparatus for producing sounds," U. K. Patent 214,281. 1922.

Paget, R. A. S. 1930. *Human Speech*. Harcourt, Brace and Company, New York, NY, USA.

Papcun, J., Hochberg, T. R., Thomas, F., Larouche, J., and Levy, S., 1992. "Inferring articulation and recognizing gestures from acoustics with a neural network trained on x-ray microbeam data," *Journal of the Acoustical Society of America*, 92:688–700.

Parke, F. I. and Waters, K. 1996. *Computer Facial Animation*. A. K. Peters, Wellesley, MA, USA.

Parlangeau, N., André-Obrecht, R., and Marchal, A. 1996. "Automatic articulatory annotation of multi-sensor speech database," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 829–832. IEEE.

Perkell, J. S., Cohen, M. H., Svirsky, M. A., Matthies, M. L., Garabieta, I., and Jackson, M. T., 1992. "Electromagnetic midsagittal articulometer systems for transducing speech articulatory movements," *Journal of the Acoustical Society of America*, 92(6):3078–3096.

Petajan, E. D. and Graf, H. P. 1996. "Robust face feature analysis for automatic speechreading and character animation," in *Speechreading by Humans and Machines: Models, Systems, and Applications*, Stork, D. G. and Hennecke, M. E., editors, pages 425–436. Springer-Verlag.

Petitto, L. A. and Marentette, P. E., 1991. "Babbling in the manual mode: Evidence for the ontogeny of language," *Science*, 251:1493–1496.

Pierce, J. R., 1969. "Whither speech recognition?," *Journal of the Acoustical Society of America*, 46(4, pt. 2):1049–1051.

Pike, K. L. 1943. *Phonetics*. University of Michigan Press, Ann Arbor, MI, USA.

Pinkerton, R., 1956. "Information theory and melody," *Scientific American*, 194:77–86.

Platt, J. 1989. *Constraint Methods for Neural Networks and Computer Graphics*. PhD thesis, California Institute of Technology, Pasadena, CA, USA.

Polansky, L., Rosenboom, D., and Burk, P. 1987. "Hierarchical Music Specification Language (HMSL): Overview (version 3.1) and notes on intelligent instrument design," in *Proceedings of the 1987 International Computer Music Conference*, Beauchamp, J., editor, pages 220–227, San Francisco, CA, USA. International Computer Music Association.

Rabiner, L. and Juang, B.-H. 1993. *Fundamentals of Speech Recognition*. Prentice-Hall, Englewood Cliffs, NJ, USA.

Rae, J. 1862 Three articles from *The Polynesian*, Sept. 27, Oct. 4, and Oct. 11. Honolulu. Reprinted in a book (Paget 1930). 1862.

Rahim, M. G. 1994. *Artificial Neural Networks for Speech Analysis/Synthesis*. Chapman and Hill, London, England, UK.

Raymond, E. S., editor. 1993. *The New Hacker's Dictionary*. MIT Press, second edition.

Robert-Ribes, J., Piquemal, M., Schwartz, J.-L., and Escudier, P. 1996. "Exploiting sensor fusion architectures and stimuli complementarity in AV [audiovisual] speech recognition," in *Speechreading by Humans and Machines: Models, Systems, and Applications*, Stork, D. G. and Hennecke, M. E., editors, pages 193–210. Springer-Verlag.

Robinson, A. J., 1994. "An application of recurrent nets to phone probability estimation," *IEEE Transactions on Neural Networks*, 5(2):298–305.

Roweis, S. T. 1999. *Data Driven Production Models for Speech Processing*. PhD thesis, California Institute of Technology, Pasadena, CA, USA.

Roweis, S. T. 2000. "Constrained hidden Markov models," in *Advances in Neural Information Processing Systems 12*.

Sakamoto, K. and Yamaguchi, K. 1992. "Recognition apparatus using articulation positions for recognizing a voice," U. S. Patent 5,175,793. 1992.

Sams, M. and Levänen, S. 1996. "Where and when are the heard and seen speech integrated: Magnetoencephalographical (MEG) studies," in *Speechreading by Humans and Machines: Models, Systems, and Applications*, Stork, D. G. and Hennecke, M. E., editors, pages 233–238. Springer-Verlag.

Sankoff, D. and Kruskal, J. B., editors. 1983. *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison*. Addison-Wesley, Reading, MA, USA.

Schroeter, J. and Sondhi, M., 1994. "Techniques for estimating vocal tract shapes from the speech signal," *IEEE Transactions on Speech and Audio Processing*, 2(1):133–150. Part Two.

Sereno, J. A., Baum, S. R., Marean, G. C., and Lieberman, P., 1987. "Acoustic analyses and perceptual data on anticipatory labial coarticulation in adults and children," *Journal of the Acoustical Society of America*, 81(2):512–519.

Shannon, C. E., 1951. "Prediction and entropy of printed English," *Bell System Technical Journal*, pages 50–64.

Shirai, K., 1993. "Estimation and generation of articulatory motion using neural networks," *Speech Communication*, 13:45–51.

Shoup, J. E. 1980. "Phonological aspects of speech recognition," in *Trends in Speech Recognition*, pages 125–138. Prentice-Hall.

Shtyrov, Y., Kujala, T., Palva, S., Ilmoniemi, R. J., and Näätänen, R., 2000. "Discrimination of speech and complex nonspeech sounds of different temporal structure in the left and right cerebral hemispheres," *Neuroimage*, 12:657–663.

Sitaram, R. N. V. and Sreenivas, T., 1997. "Incorporating phonetic properties in hidden Markov models for speech recognition," *Journal of the Acoustical Society of America*, 102(2):1149–1158. Part One.

Slager, R. P. 1993. "Apparatus for generating from an audio signal a moving visual lip image from which a speech content of the signal can be comprehended by a lipreader," U. S. Patent 5,313,522. 1993.

Sorokin, V. N., 1992. "Determination of vocal tract shape for vowels," *Speech Communication*, 11(1):71–85.

Sorokin, V. N., 1994. "Inverse problem for fricatives," *Speech Communication*, 14(3):249–262.

Sorokin, V. N. and Trushkin, A. V., 1996. "Articulatory-to-acoustic mapping for inverse problem," *Speech Communication*, 19:105–118.

Stallman, R. M. 2000. *GNU Emacs Manual.* iUniverse, Lincoln, NE, USA, thirteenth edition.

Stetson, R. H. 1988. *R. H. Stetson's Motor Phonetics: a Retrospective Edition.* College-Hill Press, Boston, MA, USA.

Stevens, S. S. and Volkmann, J., 1940. "The relation of pitch of frequency: a revised scale," *American Journal of Psychology*, 53:329–353.

Stewart, J. Q., 1922. "Electrical analog of the vocal organs," *Nature*, 110(2757):311.

Stone, M., 1990. "A three-dimensional model of tongue movement based on ultrasound and x-ray microbeam data," *Journal of the Acoustical Society of America*, 87(5):2207–2217.

Stone, M., Goldstein, M. H., and Zhang, Y., 1997. "Principal component analysis of cross sections of tongue shapes in vowel production," *Speech Communication*, pages 173–184.

Stone, M. and Lundberg, A., 1996. "Three-dimensional tongue surface shapes of English consonants and vowels," *Journal of the Acoustical Society of America*, 99(6):3728–3737.

Strange, W. and Verbrugge, R. R., 1976. "Consonant environment specifies vowel identity," *Journal of the Acoustical Society of America*, 60(1):213–224.

Ström, N. 1997. "Sparse connection and pruning in large dynamic artificial neural networks," in *Eurospeech*, volume 5, pages 2807–2810.

Sumby, W. H. and Pollack, I., 1954. "Visual contributions to speech intelligibility in noise," *Journal of the Acoustical Society of America*, 26:212–215.

Summerfield, A. Q. 1987. "Some preliminaries to a comprehensive account of audio-visual speech perception," in *Hearing by Eye: The Psychology of Lip-Reading*, Dodd, B. and Campbell, R., editors, pages 3–51. Lawrence Erlbaum Associates, Hillsdale, NJ, USA.

Tiede, M. and Vatikiotis-Bateson, E. 1994. "Extracting articulator movement parameters from a videodisc-based cineradiographic database," in *International Conference on Spoken Language Processing (ICSLP)*, volume 1, pages 44–48.

Trubetozky, N., 1939. "Grundzüge der phonologie," *Travaux du Cercle Linguistique de Prague*, 7.

Upton, H., 1968. "Wearable eyeglass speechreading aid," *American Annals of the Deaf*, 113(2):222–229.

van der Stelt, J. M. and Koopmans–van Beinum, F. 1986. "The onset of babbling related to gross motor development," in *Precursors of Early Speech*, Lindblom, B. and Zetterstrom, R., editors, pages 163–173. Academic Press.

von Kempelen, W. 1791. *Mechanismus der menschlichen Sprache nebst Beschreibung einer sprechenden Maschine; Méchanisme de la parole, suivi de la description d'une machine parlante; Mechanism of Human Speech with the Description of a Speaking Machine*. J. V. Degen, Vienna. German edition reprinted 1970 by Frommann-Holzboog, Stuttgart.

Welsh, W. J., Simons, A. D., Hutchinson, R. A., and Searby, S. 1990. "A speech-driven 'talking head' in real time," in *Proceedings of the Picture Coding Symposium*, pages 7.6.1–7.6.2, Cambridge, MA, USA.

Westbury, J. R., 1991. "The significance and measurement of head position during speech production experiments using the x-ray microbeam system," *Journal of the Acoustical Society of America*, 89:1782–1791.

Westbury, J. R. and Hashi, M., 1997. "Lip-pellet positions during vowels and labial consonants," *Journal of Phonetics*, 25(4):405–419.

Westbury, J. R., Turner, G., and Dembowski, J. 1994. "X-ray microbeam speech production database user's handbook: Version 1.0," Technical report, Univeristy of Wisconsin, Madison, WI, USA.

Wrench, A. A. 2000. "Multichannel/multispeaker articulatory database for continuous speech recognition research," in *Phonetics and Phonology in ASR*, Saarbrücken, Germany. Institute of Phonetics.

Wrench, A. A. and Richmond, K. 2000. "Continuous speech recognition using articulatory data," in *Proceedings of the International Conference on Spoken Language Processing*, Beijing, China.

Yang, W. J., Lee, J. C., Wang, H. C., and Chang, Y. C., 1988. "Hidden Markov model for Mandarin lexical tone recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-36:988–992.

Yehia, H. and Itakura, F., 1996. "A method to combine acoustic and morphological constraints in the speech production inverse problem," *Speech Communication*, 18:151–174.

Young, S., Odell, J., Ollason, D., Valtchev, V., and Woodland, P. 1997. *The HTK Book (for HTK Version 2.1)*. The Entropic Group and Cambridge University. http://htk.eng.cam.ac.uk/.

Yu, G., Russell, W., Schwartz, R., and Makhoul, J. 1990. "Discriminant analysis and supervised vector quantization for continuous speech recognition," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 685–688, Alburquerque, NM, USA.

Yuhas, B., Goldstein, M., and Sejnowski, T., 1989. "Integration of acoustic and visual speech using neural networks," *IEEE Communications Magazine*, pages 65–71.

Zacks, J. and Thomas, T. R., 1994. "A new neural network for articulatory speech recognition and its application to vowel identification," *Computer Speech and Language*, 8(3):189–209.

Zlokarnik, I., Hodgen, J., Nix, D., and Papcun, G. 1995. "Using articulatory measurements in automatic speech recognition and in speech displays for hearing impaired," in *ACCOR Workshop on Articulatory Databases*.

Zue, V. W., 1985. "The use of speech knowledge in automatic speech recognition," *Proceedings of the IEEE*, 73(11):1602–1615.

Zue, V. W. and Glass, J. R., 2000. "Conversational interfaces: Advances and challenges," *Proceedings of the IEEE*, 88(8):1166–1180.

Glossary

Acoustic model The low-level models used in speech recognition, which do not include a representation of sentence structure—the latter being part of the *grammar model.* Typically, acoustic models are *hidden Markov models.*

Acoustic recognizer In this thesis, refers to a recognizer presented with sound but not video or articulatory measurements.

Alignment, monophone/triphone A transcript of speech, composed of low-level units (monophones or triphones) and start and end times for the intervals of the input which correspond to each symbol.

Alignment, string-to-string Finding a correspondence between the letters in two strings which minimizes the edit distance (e.g., *phone error rate*) between them.

Allophones Acoustically distinct versions of a phoneme which are not semantically distinct. For example, the English phonemes /r/ and /l/ are allophones of a single phoneme in Japanese.

Articulation The movement of body parts, such as the tongue, to produce the sounds of speech.

Articulator A part of the body, such as the tongue, which is moved during speech and whose distinct motions produce distinct sounds of speech.

Automatic speech recognition *Speech recognition* by computer or other machine.

Back end The later stages of pipelined processing; in speech recognition, a typical back end is a set of *hidden Markov models.*

Bigram A pair of speech units (e.g., phones or words), especially as used for estimating symbol-to-symbol transition probabilities; purists criticize the term because it was coined by mixing Latin *bi-* with Greek *-gram.* See also *trigram* and *unigram.*

Cepstrum A signal-processing technique used as a *front end* for speech recognition; coined from "spectrum" by reversing the first syllable (Bogert et al. 1963), and thus idiosyncratically pronounced with an initial /k/. The cepstrum is the inverse Fourier transform of the logarithm of the power spectrum of a signal (section 1.10.2). Its first few coefficients contain information about the filter resonances (formants) of the vocal tract, with source activity of the vocal cords (in vowels) or another constriction (in fricatives) factored out figure 1.14. The complex cepstrum substitutes an ordinary Fourier transform for the power spectrum. Linear time-invariant filtering is equivalent to adding the complex cepstra of a convolution kernel and an input signal.

Closure In phonetics, a *stop consonant* in which airflow is blocked by holding two articulators in contact. Depending on context, /d/ in English may be either a closure or a flap.

Digram Synonym for *bigram*.

Digraph A sequence of two letters producing a single sound, such as "th;" dates to 1788 (OUP 1989).

Feature set In pattern recognition (Duda and Hart 1973), the parameters such as coefficients of the *cepstrum* that are passed from a *front end* to a *back end* in pipelined speech recognition; in linguistics, a set of discrete characteristics which distinguish phone-like units (Jakobson et al. 1952) (Liberman et al. 1967).

Fenones Data-driven substitutes for *phones* and *phonemes* in representing pronunciation. The fenone representation of an acoustic input is the output symbol sequence of a discrete front end such as vector quantization. The word was coined, according to a first-hand source, by combining "F.E.-" for "front end" with "none," the latter "intended to lend the term scientific respectability" (Jelinek 1998a).

Flap In phonetics, a *stop consonant* produced by briefly flapping one articulator against another. Depending on context, /d/ in English may be either a closure or a flap.

Fricative A consonant such as /f/ in which the tongue is held at a point of constriction close to another articulator.

Front end The early stages of pipelined processing; in speech recognition, a typical front end includes the *cepstrum* and numerical differentiation.

Gesture In articulatory phonology and the *gesture theory*, refers to a motion which is part of the production of speech (Browman and Goldstein 1989).

Gesture theory A hypothesis about the origins of speech and nature of its perception: that speech originated in gestures (Rae 1862) and that "we lip-read by ear" (Paget 1930).

Grammar model A means for determining which symbol sequences are valid outputs of the recognizer. More generally, a scheme for estimating the a priori probability of different output sequences (e.g., n-*gram language model*).

Han'gŭl A Korean alphabet whose consonant letter forms are based on the shape of the articulators critical to its pronunciation.

Heteroscedastic Having different variances or nonuniform variance.

Hidden Markov model (HMM) A probabilistic state machine used to model the process of speech, typically used as a *back end* for *automatic speech recognition*.

Hidden Markov model toolkit (HTK) A commercial product which implements many algorithms for *automatic speech recognition*, including the cepstrum and continuous-emission-density HMMs.

Inverse problem In speech recognition, refers to speech production by humans or synthesis by machine.

Kinematics Geometric description of motion, without reference to mass or force.

Labiodental A sound such as /f/ made with the lips and teeth (1669) (OUP 1989).

Language model Commonly used term for *grammar model*, as contrasted with *acoustic model*. The term is avoided in this thesis since acoustic phonetics are a part of language as well.

Lipreading Observing the face to recognize speech, with or without acoustic information; appears in print in 1874. Also called labiomancy, from 1686 (OUP 1989).

Mixture In speech recognition, confusingly used to refer to a Gaussian in a weighted-sum mixture of Gaussians (Rabiner and Juang 1993) (Young et al. 1997); in this thesis, mixture always refers to the sum itself, not its constituent terms. Unlike radio engineering, in which a mixer combines signals nonlinearly, mixtures in speech recognition are linear combinations of probability densities.

Monophone A single unit of speech used in a recognizer. Typically, and in this thesis, monophones are defined by engineering expediency, and may not meet the strict linguistic definitions of an allophone or a phoneme. A monophone recognizer is context independent in the sense that the monophone models do not change based on surrounding sounds—a *triphone* recognizer is context dependent.

Motor space A set of possible articulator states.

Motor theory The idea that "the objects of speech perception are the intended phonetic gestures of the speaker, represented in the brain as invariant motor commands... for example, 'tongue backing,' 'lip rounding,' and 'jaw raising'..." (Liberman and Mattingly 1985).

North inland cities shift A reorganization of the short vowels of English (e.g., "black" for "block" and "boss" for "bus"), displayed by some urban speakers from parts of the United States, including Wisconsin, where the data of chapter 3 were taken.

Phone error rate (PER) Given a target sequence τ of phone-like units which were spoken and a recognizer's guessed sequence γ, in this thesis the phone error rate E is defined as

$$E = \frac{D(\tau,\gamma) + S(\tau,\gamma) + I(\tau,\gamma)}{\|\tau\|}$$

where $\|\tau\|$ is the number of symbols in τ; and $D(\tau,\gamma)$, $S(\tau,\gamma)$, and $I(\tau,\gamma)$ are the number of deletions, substitutions, and insertions which convert τ into γ, when the operations are chosen so as to minimize E. The numerator is a type of *Levenshtein distance* (Levenshtein 1966).

Phone A minimal unit of speech defined acoustically and in a way not tied to any particular language, in contrast to a *phoneme*.

Phoneme A minimal unit of speech that distinguishes words in a specific language. See also *allophone* and *phone*.

Phone-like unit (PLU) A unit of speech used in a recognizer which is similar to a phoneme, and is expedient for engineering but may not meet precise linguistic criteria.

Quefrency The independent variable of the domain of the *cepstrum*, just as frequency is the independent variable of the domain of the power spectrum. Quefrency is measured in units of time. Small quefrencies correspond to coarse spectral shape; they are retained for automatic speech recognition, while cepstral coefficients at larger quefrencies are discarded.

Speechreading A term for visual speech perception reflecting the fact that one watches more than just the lips. The common word *lipreading*, which dates back 127 years, is used in this thesis because it is generally understood to include the observation of both "lip and facial movements" (Merriam-Webster 1996).

Speech recognition A type of voice recognition problem in which the goal is to determine what was said, rather than who was speaking.

Stop (consonant) A sound of speech such as /t/ for which the tongue makes an abrupt transition towards or away from some other articulator.

Trigram A general term for three units of speech in a row. In speech recognition, it typically refers to calculation of third-order statistics: symbol probabilities based on the two preceding symbols (Jelinek 1998b). In reference specifically to letter sequences, the word dates back to 1606 (OUP 1989).

Triphone A unit of speech consisting of three phones in a row: previous, current, and following. The previous and following phones may not be specifically identified, but instead described as belonging to a particular class. Such generalized triphones are used in state-of-the-art recognizers.

Unigram A single unit of speech, for which a prior probability is estimated without considering transition probabilities; used only in contrast to *bigram* and *trigram* language models. The word is a combination of the Latin *uni-* with the Greek *-gram*; the Greek prefix for "one" is *mono-*, but "monogram" commonly refers to "two or more letters interwoven together" (OUP 1989)—i.e., a bigram or trigram.