

**FEATURES AND MEASURES FOR
SPEAKER RECOGNITION**

By

JOSEPH PAUL CAMPBELL, JR.

**Bachelor of Science
Rensselaer Polytechnic Institute
Troy, New York
1979**

**Master of Science
The Johns Hopkins University
Baltimore, Maryland
1986**

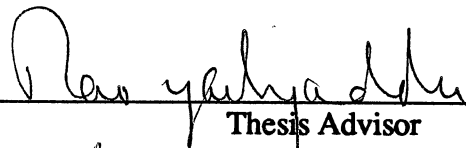
**Submitted to the Faculty of the
Graduate College of the
Oklahoma State University
in partial fulfillment of
the requirements for
the Degree of
DOCTOR OF PHILOSOPHY
December, 1992**

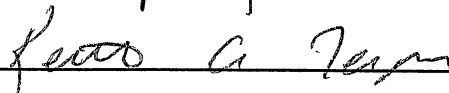
1992D C188f

Thesis
1992D
C188f

FEATURES AND MEASURES FOR
SPEAKER RECOGNITION

Thesis Approved:

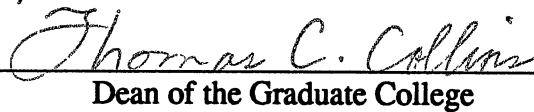

Thesis Advisor










Dean of the Graduate College

PREFACE

This work derives and demonstrates new and powerful features and measures for automatic speaker recognition and compares them with traditional ones using speaker discrimination criterion. Automatic speaker recognition is the use of a machine to recognize a person from a spoken phrase. Speaker recognition systems can be used in two modes: to *identify* a particular person or to *verify* a person's claimed identity.

New perceptually based features were found which, unfortunately, did not outperform traditional speech production features with respect to speaker identification errors. Powerful new production features and measures for speaker verification were discovered. The main contribution of this work is a new information-theoretic shape measure between line spectrum pair frequency features. I call this new measure the *divergence shape* because it can be interpreted geometrically as the shape of an information-theoretic measure called divergence. LSPs were found to be very effective features in this divergence shape measure. Experimental results show this combination yields 99.95% correct speaker identification. The corresponding 0.05% speaker identification error is superior to the performance of any other claim reported in the literature by over an order of magnitude.

As automatic speaker authentication systems gain widespread use, it is imperative to understand the errors made by these systems. There are two types of errors: the false acceptance of an invalid user (type I) and the false rejection of a valid user (type II). It takes two people to make a false acceptance error: an impostor and a target. Because of this hunter and prey relationship, the impostor is referred to as a *wolf* and the target as a *sheep*. Although automatic voice verification is not new, specific understanding of and a

means to reduce false acceptance errors have been virtually ignored in the literature. False acceptance errors are the ultimate concern of high-security speaker authentication applications. This dissertation develops a method to reduce false acceptance errors due to wolves and sheep.

I'm grateful to the U.S. Government for awarding me a fellowship to pursue these studies and to Ken Rose, John Lee, Ron Benincasa, and my coworkers for their support. Our senior speech scientist, Tom Tremain, has been a guiding light throughout my career and this research. I'm grateful to all the wonderful teachers and coaches I've had along the way, especially Joe Basalyga, Bart Bonney, Jess Fussell, John Miller, Don Starkweather, and my father. I wish to thank Bishnu Atal, Tom Crystal, Rich Dean, John Endsley, Alan Higgins, Fred Juang, Dave Murley, and Frank Soong for many fruitful discussions. I'm grateful for the late night assistance given by Bill Budney, Dave Kemp, Jerry Lathem, Craig Reese, and Vanoy Welch to help keep the computers running. I wish to acknowledge contributions from Brian Evans, Dik Hermes, Luc Van Immerseel, Steve Leon, Jean-Pierre Martens, Tony Richardson, and Malcolm Slaney.

The kindness, patience, encouragement, inspiration, scholarship, and knowledge of my major advisor, Professor Rao Yarlagadda, are extraordinary. My fellowship award was for any school of my choosing and I was fortunate to have spent it at OSU under Prof. Yarlagadda. I'm grateful for the support of my advisory committee: Professors Baker, Rhoten, Pentz, and Teague. I'll fondly remember Prof. Baker's granddaughter with my daughters at the Stillwater playground, listening to Dvořák with Prof. Rhoten, enjoyable conversations with Prof. Pentz after every class, and OSU football games with Prof. Teague and his wife Sherry. Barbara Caldwell's efficiency and leg work allowed me to meet the deadlines. The people of the OSU campus are among the warmest I've known. I'll miss the many faculty and student friends I made there, including members of the HP-48 User Group, the Macintosh User Groups, and the W5YJ Amateur Radio Club.

I'm especially grateful to my family for their support. The perseverance my parents instilled in me allowed me to endure this effort. I'm grateful to my uncle, Bill Campbell, for fostering an appreciation of science, mathematics, and tennis at an early age. My wife Shawn provided love and editing and interrupted her technical writing career to join me with our daughters on this journey. My daughters, Elizabeth and Emily, gave me the motivation to finish this work in a timely fashion. I look forward to giving my family back the time they've given me to pursue this research. I regret that some of my family didn't live to see the completion of this effort. I dedicate this work to them and to the rest of my family.

TABLE OF CONTENTS

Chapter	Page
I. INTRODUCTION.....	1
Motivation.....	3
Problem Formulation	4
Generic Voice Verification	5
Overview of Dissertation	6
Previous Work	6
Proposition	9
Experimental Results and Observations	10
Discussion	11
II. SPEECH PROCESSING.....	13
Voice Signal Acquisition	13
YOHO Database	14
Speech Production	17
Linear Prediction.....	20
Reflection Coefficients	24
Log Area Ratios	25
Arcsin Reflection Coefficients.....	27
Line Spectrum Pair Frequencies	27
Speech Perception.....	29
Perceptually Motivated Features.....	32
Pitch	34
Auditory Model Pitch Estimation	36
Loudness	38
Perceptual-Model Filterbank	40
III. LP AND SINGULAR VALUE DECOMPOSITION	43
Definition of SVD.....	45
SVD-Based Speech Models.....	50
Matrix Form of Linear Prediction.....	51
Linear Prediction Properties	52
SVD of LP Impulse Response Matrix	52
SVD Transform Representation.....	54
Perceptually Based SVD.....	56
Speech Coding	58
Vector Quantization Code Book Design.....	59
Fast Vector Quantization	59

Chapter	Page
Speaker Authentication.....	60
SVD Advantages.....	60
Generalized SVD	61
IV. FEATURE SELECTION AND MEASURES	63
Traditional Feature Selection.....	63
Normal Density With Equal Means.....	67
Importance Sampling.....	70
Mean and Covariance Estimation	70
Divergence Measure	73
Equal Covariance Divergence.....	76
Equal Mean Divergence.....	77
Divergence Properties.....	78
Example of Equal Mean Divergence	80
Bhattacharyya Distance	83
V. PATTERN MATCHING	85
Template Models	86
Dynamic Time Warping	86
Vector Quantization Source Modeling	88
Nearest Neighbors.....	88
Stochastic Models	90
VI. CLASSIFICATION AND DECISION THEORY	93
Hypothesis Testing.....	93
Receiver Operating Curve.....	96
Data Fusion	97
VII. PERFORMANCE	99
DTW System.....	99
ROC of DTW and NN Systems.....	100
Wolves and Sheep.....	102
LSP Divergence Shape Speaker Identification.....	110
VIII. INNOVATIONS	115
Accomplishments.....	119
IX. SUMMARY AND CONCLUSIONS.....	120
The Problem.....	120
Important Findings.....	120
Suggestions for Future Research or Study.....	121

Chapter	Page
CITATIONS.....	123
BIBLIOGRAPHY.....	129

LIST OF TABLES

Table	Page
I-1. Sources of Verification Error	4
I-2. Selected Chronology of Speaker Recognition Progress.....	8
II-1. Frequency Response of YOHO Decimation Filter	15
II-2. The YOHO Database	16
II-3. Example Linear Predictor Coefficients	28
II-4. Common Speech Processing Features.....	33
II-5. Acoustic and Perceptual Correlates.....	34
II-6. Perceptual-Model Filterbank.....	41
III-1. Traditional Versus Modern Methods	44
III-2. Fundamental Subspaces	47
III-3. Calculating the Masking Threshold	58
VI-1. Probability Terms and Definitions	95
VII-1. Known Wolves and Sheep, DTW System.....	100
VII-2. Wolf and Sheep Sexual Relationships.....	101
VII-3. Errors of Various Features and Measures	114
VIII-1. Innovations.....	116
VIII-2. Relative Performance.....	117

LIST OF FIGURES

Figure	Page
I-1. Speech Processing	1
I-2. Typical Speaker Authentication Setup.....	3
I-3. Generic Speaker Verification System	5
II-1. Human Vocal System.....	18
II-2. Acoustic Tube Model of Speech Production.....	26
II-3. Frequency Response.....	29
II-4. LSP Frequencies and LP Poles in the z-Plane.....	30
II-5. The Human Ear	31
II-6. The mel Pitch Scale.....	35
II-7. The Bark Masking Scale	36
II-8. Auditory Model Pitch Extractor.....	37
II-9. Equal-Loudness Level Contours	39
II-10. Perceived Loudness Scale	40
II-11. Perceptual-Model Filterbank.....	42
III-1. LP Synthesis.....	50
III-2. Sinusoidal Structure of Right SVs of H.....	56
III-3. Narrow-Band Structure of Power Spectrum of Right SVs of H	57
IV-1. Linear Transformation.....	66
IV-2. Unequal Covariance	68
IV-3. A Bimodal Class.....	69
IV-4. LSP Covariance Matrices: Different Sessions, Same Speaker.....	72

Figure	Page
IV-5. LSP Covariance Matrices, Different Speakers.....	72
IV-6. Original Observation Vectors.....	81
V-1. Dynamic Time Warping Two Energy Signals	87
V-2. Nearest Neighbor Method	89
V-3. An Example of a Three-State Hidden Markov Model	91
VI-1. Valid and Impostor Densities	94
VI-2. An Example of Score Densities.....	95
VI-3. Hypothetical Receiver Operating Curves.....	97
VII-1. Receiver Operating Curves	102
VII-2. Speaker vs FA Errors for DTW System's Wolves and Sheep	103
VII-3. FA Errors for DTW System's Wolves and Sheep.....	104
VII-4. Speaker vs FA Errors for NN System's Wolves and Sheep.....	105
VII-5. Speaker vs FA Errors for DTW and NN Systems' Sheep.....	106
VII-6. Speaker vs FA Errors for DTW and NN Systems' Wolves	107
VII-7. FA Errors vs Session Number for NN System.....	108
VII-8. Wolf and Sheep Pairings of the DTW System.....	109
VII-9. LSP Divergence Shape (1 error)	111
VII-10. LSP Bhattacharyya Shape (2 errors)	112
VII-11. LSP Bhattacharyya Distance (4 errors)	112
VII-12. LSP Divergence Measure (3 errors).....	113
VIII-1. Signal Processing Blocks of New System	118

NOMENCLATURE

\mathbf{A}	matrix
$ \mathbf{A} $	determinant
$\ \mathbf{A}\ $	norm
\mathbf{A}^T	transpose
\mathbf{A}^{-1}	inverse
$\mathbf{A}^{(-1)}$	pseudoinverse
a	scalar
\mathbf{a}	vector (all vectors are columnwise, i.e., $\mathbf{a} \in \mathfrak{R}^{m \times 1}$)
$E[\cdot]$	statistical expectation
e	base of natural logarithms (2.71828...)
$\ln(\cdot)$	logarithm base e
N	scalar (e.g., summation limit)
$\sim N(\boldsymbol{\mu}, \mathbf{C})$	normal density with mean vector $\boldsymbol{\mu}$ and covariance matrix \mathbf{C} : $p(\mathbf{x}) = (2\pi)^{-n/2} \mathbf{C} ^{-1/2} \exp\left[-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \mathbf{C}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right]$ (normal pdf)
\mathfrak{R}	real space
\mathfrak{R}^m	m dimensional real space
$\mathfrak{R}^{m \times n}$	$m \times n$ dimensional real space
$\text{tr}[\mathbf{A}]$	trace
\exists	there exists
\forall	for all
\vee	for
\in	membership

ACRONYMS

A/D	analog-to-digital
AGN	additive Gaussian noise
AMPEX	auditory model-based pitch extractor
ANOVA	analysis of variance
AR	autoregressive
BPF	band-pass filter
DTW	dynamic time warping
EER	equal error rate
EVD	eigenvalue-eigenvector decomposition
FA	false acceptance (type I error)
FFT	fast Fourier transform
FIR	finite duration impulse response
FR	false rejection (type II error)
GSVD	generalized singular value decomposition
HMM	hidden Markov model
iff	if and only if
i.i.d.	independent identically distributed
JND	just noticeable difference
KLE	Karhunen-Loève expansion
LAR	log-area ratio
LP	linear predictor (or linear prediction)
LPC	linear predictive coding (or linear prediction coefficient)

LSP	line spectrum pair
MLE	maximum likelihood estimate
MSE	mean squared error
NN	nearest neighbor
PC	prediction coefficient
pdf	probability density function
RC	reflection coefficient
ROC	receiver operating curve
SHS	subharmonic summation
SNR	signal-to-noise ratio
SV	singular vector
SVD	singular value decomposition
UBE	unbiased estimate
VQ	vector quantization
ZIR	zero-input response
ZSR	zero-state response

CHAPTER I

INTRODUCTION

Speech processing is a diverse field with many applications. Figure I-1 shows a few of these areas and how the topic of this research, shown in the box, relates to the rest of the field.

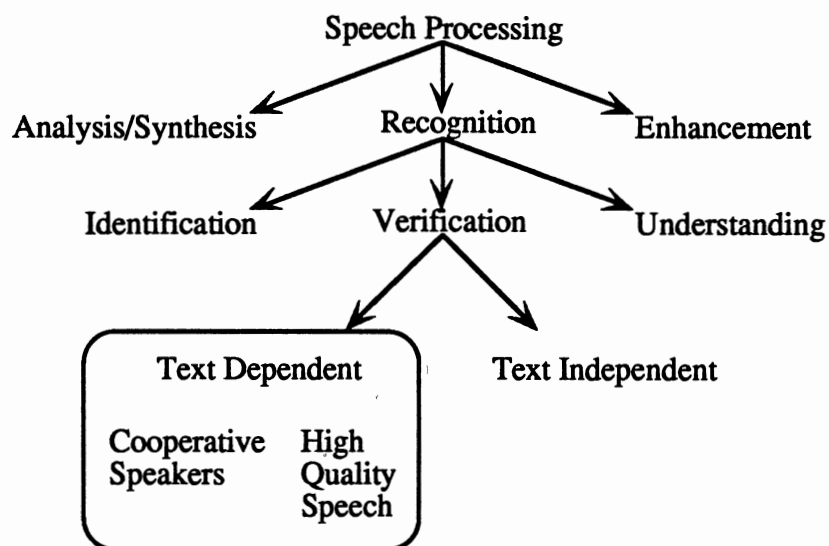


Figure I-1. Speech Processing

Automatic speaker authentication is the use of a machine to authenticate a person's claimed identity from his voice. The literature abounds with different terms for speaker authentication, including speaker verification, voice authentication, voice verification,

talker authentication, and talker verification. Additionally, the term recognition encompasses verification and identification. General reviews of speaker recognition are given in the Citations (Atal 1976; Doddington 1985; Furui 1991; O'Shaughnessy 1987; Rosenberg 1976; Rosenberg and Soong 1992; Sutherland and Jack 1988).

Speaker authentication is defined as deciding if a speaker is whom he claims to be. This is different than the speaker identification problem, which is deciding if a speaker is a specific person or is among a group of persons. In speaker authentication, a person makes an identity claim (e.g., entering an employee number or presenting a smart card). In text-dependent verification, the system then prompts the claimant (visually or orally) to say a phrase. The claimant speaks the phrase into a microphone. This signal is analyzed by an authentication system that makes the binary decision to accept or reject the user or it may request additional input before making the decision.

A typical automatic speaker authentication setup is shown in Figure I-2. The claimant has previously enrolled in the system and he presents an encrypted smart card containing his identification information. He then attempts to be authenticated by speaking a prompted phrase(s) into the microphone. There is generally a tradeoff between the test session duration and accuracy. In addition to his voice, ambient room noise and delayed versions of his voice enter the microphone via reflective acoustic surfaces. Prior to an authentication session, users must enroll in the system (typically under supervised conditions). During this enrollment, voice models are generated and stored (possibly on a smart card) for use in later authentication sessions. There is generally a tradeoff between accuracy and the duration and number of enrollment sessions.

Many factors can contribute to verification errors. Table I-1 lists some of the human and environmental factors that contribute to authentication errors, some of which are shown in Figure I-2.

These factors are generally outside the scope of algorithms or are better corrected by means other than algorithms and, therefore, they will not be discussed further. However,

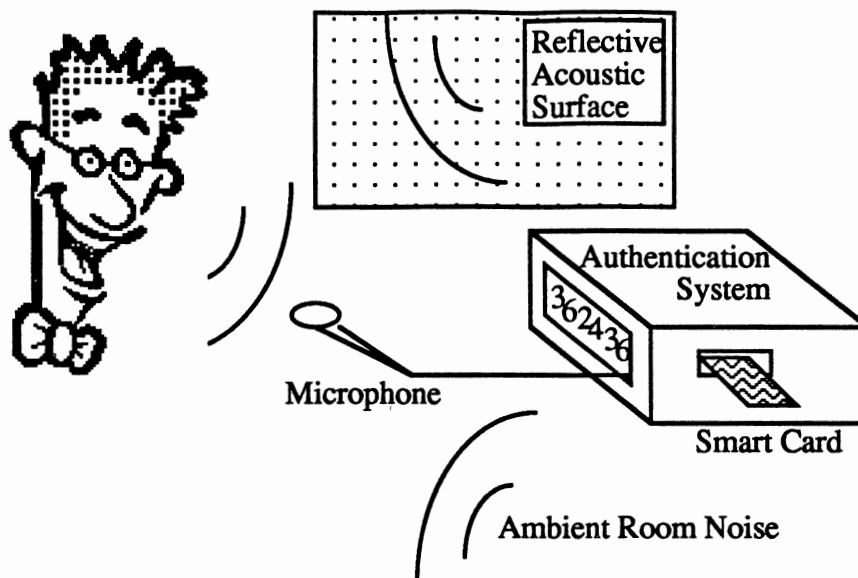


Figure I-2. Typical Speaker Authentication Setup

these factors are important because, no matter how good a speaker authentication algorithm is, human error ultimately limits its performance.

Motivation

We are now living in an information age. Information has become a valuable commodity and needs to be protected. Institutions make critical decisions based on the information they have. Adversaries often try to acquire another institution's sensitive information to gain an unfair competitive or strategic advantage. Therefore, sensitive information must be protected. A means to this end is the U.S. Government's secure voice program. Under Presidential Directive 24, the goal of this program is to field 1 million secure voice/data terminals within the next few years.

The widespread proliferation of secure voice equipment lacking user verification capability increases the potential for their abuse. Speaker authentication is perhaps the most natural method to solve the problems of unauthorized use and multilevel access

TABLE I-1
SOURCES OF VERIFICATION ERROR

Misread or misspoken prompted phrases
Time varying (intra- or intersession) microphone placement
Poor or inconsistent room acoustics (e.g., multipath and noise)

control. Past speaker authentication research has almost totally ignored verification errors specifically due to false acceptance of impostors. Research in this area is required before speaker authentication systems can be trusted to guard against type I errors.

Unlike other personal authentication methods, your voice cannot be lost or forgotten and, furthermore, speaker verification systems can be made resilient to attack from mimicry by humans and tape recorders.

Problem Formulation

Speech is a complicated signal produced as a result of several transformations occurring at several different levels: semantic, linguistic, articulatory, and acoustic. Differences in these transformations appear as differences in the acoustic properties of the speech signal. Speaker-related differences are a result of a combination of anatomical differences inherent in the vocal tract and the learned speaking habits of different individuals. In speaker authentication, both these differences can be used to discriminate between speakers.

The focus of this research is on understanding the causes and reducing type I speaker authentication errors without raising type II errors to unacceptable levels. Past speaker authentication research has almost totally ignored those errors specifically caused by false

acceptance of impostors. Research in this area is required before authentication systems can be trusted.

Generic Voice Verification

The general approach to voice verification consists of five steps: digital speech data acquisition, feature extraction, pattern matching, making an accept/reject decision, and enrollment to generate speaker reference models. A block diagram of this procedure is shown in Figure I-3. Feature extraction maps each interval of speech to a multidimensional feature space. (A speech interval typically represents 20 ms of the speech waveform and is referred to as a frame of speech.) This sequence of feature vectors, x_i , is then compared to speaker models by pattern matching. This results in a match score, z_i , for each vector or sequence of vectors. The match score measures the similarity of the computed input feature vectors to models of the claimed speaker or feature vector patterns for the claimed speaker. Finally, a decision is made to either accept or reject the claimant according to the match score or sequence match scores, which is a hypothesis testing problem.

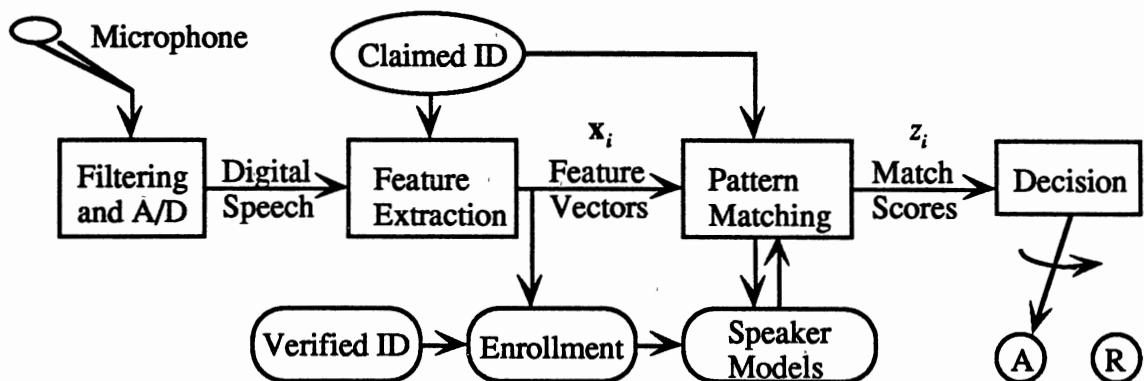


Figure I-3. Generic Speaker Verification System

For speaker verification, features that exhibit high interspeaker variability and low intraspeaker variability are desired. The pattern matching approach depends mainly upon the type of model being used. The model can be template or stochastic based.

Overview of Dissertation

This dissertation is comprised of nine chapters. The purpose of this introductory chapter is to present some general motivational framework for speaker recognition, an overview of the entire dissertation, a discussion of the previous work in the area of speaker recognition, and a discussion of the contributions of the author's research.

Chapter II contains an overview of digital signal acquisition, speech production, speech signal processing, linear prediction, and speech perception. Chapter III presents singular value decomposition in the context of linear prediction and speech perception. Chapter IV presents feature selection, estimation of mean and covariance, divergence, and Bhattacharyya distance. This chapter is highlighted by the development of the divergence shape measure and the Bhattacharyya distance shape. Chapter V introduces statistical pattern matching and receiver operating curves and Chapter VI presents classification and statistical decision theory. Chapter VII demonstrates the speaker identification performance of the new algorithm relative to two reference speaker verification algorithms and Chapter VIII presents the innovations of this research. Chapter IX concludes by reviewing the problem at hand, summarizing the major contributions of the research contained in this document, and by suggesting future research.

Previous Work

There is considerable speaker verification activity in industry, national laboratories, and universities. Both AT&T Bell Laboratories and Texas Instruments have researched and designed several generations of speaker verification systems. Currently, ITT,

Bellcore, Siemens, and the regional Bell operating companies are conducting research and development (Naik 1990). There are commercial offerings from Voxtron, ECCO, and Alpha Microsystems. Sandia National Laboratories and the U.S. Government are conducting evaluations of speaker authentication systems. The majority of speaker verification research at these companies is directed at verification over telephone lines. One notable exception is ITT's project YOHO.

Table I-2 shows a sampling of the chronological advancement in speaker verification. The following terms are used to define the columns in Table I-2: Source refers to a citation in the Citations, Org is the company or school where the work was done, Features are the signal measurements (e.g., cepstrum), Input is the type of input speech (laboratory, telephone, or office quality), Text indicates whether text-dependent or text-independent phrases are used, Method is the heart of the matching process, Pop is the population size (number of people), and Error is the equal error percentage for speaker verification systems or the error percentage for speaker identification systems given the specified duration of test speech in seconds. This data is presented to give a simplified general view of past speaker recognition research. It is difficult to make meaningful comparisons between the text-dependent and the generally more difficult text-independent tasks. It is also difficult to compare between the binary-choice verification task and the generally more difficult multiple-choice identification task (Doddington 1985).

The performance of current systems makes them suitable for many practical applications. However, for high-security applications, these performance levels would generally be unacceptable; they would need to be used in combination with other authenticators (e.g., smart card). The level of performance achieved in this work is acceptable for many high-security applications.

TABLE I-2
SELECTED CHRONOLOGY OF SPEAKER RECOGNITION PROGRESS

Source	Org	Features	Input	Text	Method	Pop	Error
(Atal 1974)	AT&T	Cep	Lab	Dep	Pattern Match	10	2%@0.6s
(Markel and Davis 1979)	STI	LPC	Lab	Indep	Long Term Statistics	17	2%@39s
(Furui 1981)	AT&T	Normalized Cep	Phone	Dep	Pattern Match	10	0.2%@3s
(Schwartz and others 1982)	BBN	LAR	Phone	Indep	Non-parametric pdf	21	3%@2s
(Li and Wrench 1983)	ITT	LPC, Cep	Lab	Indep	Pattern Match	11	21%@3s 4%@10s
(Doddington 1985)	TI	Filter-bank	Lab	Dep	DTW	200	~0.8%@6s
(Higgins and Wohlford 1986)	ITT	Cep	Lab	Indep	VQ	10	10%@2.5s 5%@10.5s
(Soong and others 1987)	AT&T	LPC	Phone	Dep (digits)	VQ	100	6%@1s 1.5%@5s
(Attili and others 1988)	RPI	Cep, LPC, Autocorr	Lab	Indep	Projected Long Term Statistics	90	4%@3s
(Higgins and others 1991)	ITT	LAR, LPC Cep	Office	Dep	DTW Likelihood Scoring	186	0.7%@20s
(Tishby 1991)	AT&T	LPC	Phone	Dep (digits)	HMM (mix AR)	100	5.6%@1s 0.8%@5s

Proposition

The goal of this research is to reduce false acceptance errors in speaker authentication systems. The following topics are covered: speech processing by humans and machine, pattern recognition, decision theory, and to understand and reduce false acceptance errors in speaker authentication systems.

The focus of this research is to discover powerful features and measures for automatic verification of a person's identity from a spoken phrase. The scope of this study is limited to speech collected from cooperative users in real-world office environments and without adverse microphone or channel impairments. Unlike other personal authentication methods, your voice cannot be lost or forgotten and, furthermore, speaker verification systems can be made resilient to attack from mimicry by humans and tape recorders. The success of speaker verification systems depends directly upon the power of the features and measures used to discriminate among people. Speaker verification applications include access control, telephone banking, and telephone credit cards. The LA Times recently reported that \$1.2 billion is lost annually from telephone calling card fraud and the accounting firm of Ernst and Young estimates that high-tech computer thieves in the U.S. steal \$3 to \$5 billion annually! Automatic voice verification technology can substantially reduce this crime by authenticating these fraudulent transactions. As automatic speaker authentication systems gain widespread use, it is imperative to understand the errors made by these systems. There are two types of errors: the false acceptance of an invalid user (type I) and the false rejection of a valid user (type II). It takes a pair of subjects to make a type I error: an impostor and a target. Because of this hunter and prey relationship, in this work, the impostor is referred to as a wolf and the target as a sheep. Although automatic voice verification is not new, specific understanding of and the means to reduce false acceptance errors have been virtually ignored in the literature. False acceptance errors are the ultimate concern of high-security

speaker-authentication applications. This dissertation develops a method and outlines a research plan to understand the causes of speaker authentication errors and to reduce false acceptance errors due to wolves and sheep. A thorough literature review of over 300 references was conducted. Then, concepts were synthesized from diverse fields, including signal processing, information theory, pattern recognition, physiology, and speech production and perception. After identifying over a dozen innovations, they were compared analytically as well as by computer simulation. All FORTRAN and C language computer simulations were verified using MatLab™ or Mathematica™ high-level languages. To ensure that statistically meaningful results and conclusions could be obtained, an extensive experiment was necessary. A database of 186 people collected over a 3 month period was used in the experiments. These experiments consumed over 3 months of Cray-2 supercomputer time and 5 billion bytes of storage; however, a speaker verification system using methods presented in this dissertation would be practical to implement in software on a modern personal computer. Since this exceeded the computational and storage capacity available at OSU, the experiments were performed at a Department of Defense facility.

Experimental Results and Observations

In this research, new features and measures for speaker verification were explored and compared with traditional ones using speaker discrimination criterion. It was found that new perceptually based features did not outperform traditional speech production features with respect to speaker identification errors. Also discovered were powerful new production features and measures for speaker verification. Experimental results show that these new features and measures yield 0.05% speaker identification error. This is an order of magnitude better than the performance of any other claim reported to date. The main contribution of this work is a new information-theoretic shape measure between line spectrum pair (LSP) frequency features. This new measure, the *divergence shape*, can be

interpreted geometrically as the shape of an information-theoretic measure called divergence. The LSPs were found to be very effective features in this divergence shape measure.

Discussion

The LSP divergence shape is shown to have strong speaker discriminatory power. The LSP and LP cepstral features were found to be powerful in the divergence measures and Bhattacharyya distances. Numerical limitations precluded the use of sophisticated optimum information-theoretic, linear feature selection techniques.

A speaker identification test yielded 99.95% correct speaker identification using motivated speakers with high-quality telephone-bandwidth speech collected in real-world office environments under a constrained grammar (YOHO). This experiment uses 44 people from the YOHO database with 80 seconds of speech for training and testing. Each speaker is compared to a different session of himself and to 2 sessions of 43 other speakers. The “closest” speaker to each candidate is identified. Only 1 false identification error was made on a total of 1936 tests. The “closeness” criterion yielding this result is the information-theoretic divergence measure without mean information. This outperformed divergence with means (3 errors), Bhattacharyya distance (4 errors), and Bhattacharyya distance without means (2 errors). The features yielding these results are the line spectrum pair frequencies. Using the same speech data, conventional Euclidean distance commits 38 errors (1.96% error) and conventional Mahalanobis distance makes 21 errors (1.08%). The LSP divergence shape performs the best among these tests with only 1 error (0.05%). The implication of this powerful new measure is vastly improved speaker recognition performance relative to the state of the art.

In addition to being a powerful measure, the data used by the LSP divergence shape to characterize a speaker can be compactly represented. In these experiments, each speaker is represented by the covariance matrix of his 10 LSP frequencies. A covariance

matrix can be represented by its upper (or lower) triangular section. Exploiting this symmetry, a person's 10 x 10 covariance matrix can be represented with only 55 elements.

In a practical sense, a large portion of the billions of dollars currently lost to fraud annually could be saved by verifying transactions through the application of this powerful LSP divergence shape measure to speaker verification systems.

The following chapter contains an overview of digital signal acquisition, speech production, speech signal processing, linear prediction, and speech perception.

CHAPTER II

SPEECH PROCESSING

Speech processing extracts the desired information from a speech signal. To process a signal by a digital computer, the signal must be represented in digital form so that it can be used by a digital computer.

Voice Signal Acquisition

Initially, the acoustic sound pressure wave is transformed into a digital signal suitable for voice processing. A microphone or telephone handset can be used to convert the acoustic wave into an analog signal. This analog signal is conditioned with antialiasing filtering (and possibly additional filtering to compensate for any channel impairments). The antialiasing filter limits the bandwidth of the signal to approximately the Nyquist rate (half the sampling rate) before sampling. The conditioned analog signal is then sampled to form a digital signal by an analog-to-digital (A/D) converter. A/D converters for speech applications typically generate 8,000 to 20,000 samples per second with 10 to 14 bit resolution samples. Oversampling is commonly used to allow a simpler analog antialiasing filter and to precisely control the fidelity of the sampled signal.

In local speaker authentication applications, the analog channel is simply the microphone, its cable, and analog signal conditioning. Thus, the resulting digital signal can be very high quality; as opposed to, for example, authentication using analog signals over long-distance telephone lines.

YOHO Database

This research will initially be based on high-quality signals for benign-channel speaker authentication applications. The primary database for this research is known as the YOHO database and was collected by ITT under a U.S. Government contract administered by the author. This database is already in digital form, so the first signal processing block of the verification system in Figure I-3 (signal conditioning and acquisition) is taken care of.

The signal conditioning and acquisition was designed by the author using a 4-times oversampling method to provide bandwidth and linear phase up to 3.8 kHz. First, the analog signal is low pass filtered at approximately 5 kHz by a mild 4th order elliptic analog antialiasing filter that has negligible effect on the signal below 4 kHz. This analog antialiasing filter sufficiently limits the bandwidth to 16 kHz to prevent aliasing when it is then oversampled at 32 kHz with 12 bits of precision. Next, the 32-kHz sampled signal is passed through a 255-tap, finite duration impulse response (FIR) digital bandpass filter. This digital filter limits the bandwidth of the signal so that it can be decimated by 4:1 to arrive at the final desired sampling frequency of 8 kHz. Using iterative inverse- and forward-Fourier transforms, the author designed a frequency-sampling symmetric-FIR filter-design routine to determine the 255 coefficients that best approximate a secure voice terminal's input characteristics in a least mean-square magnitude-response error sense. The resulting response models the STU-III secure voice terminal's input characteristics very closely and is given in Table II-1.

The key to oversampling is that the analog antialiasing filter need not have steep skirts in the vicinity of the half sampling frequency, as in the Nyquist sampling methods, whereas the symmetric digital FIR filter has linear phase and can have arbitrarily flat magnitude response. The advantage of the oversampling method is that the magnitude and phase distortions near the half sampling frequency are far less than is common in

TABLE II-1
 FREQUENCY RESPONSE OF YOHO
 DECIMATION FILTER

Frequency (Hz)	Response (dB)
0	-25
< 50	-21
100	-7
150	-2
200	-0.2
200 – 3600	-0.2 to +0.3 peak ripple
3600	-0.2
3800	-3
4000	-25
4400	-42
> 5000	-50
16,000	-57

traditional Nyquist sampling methods. For example, Digital Sound Corporation's -3 dB analog bandwidth for 8 kHz sampling is only 3.6 kHz, as opposed to the 3.8 kHz, -3 dB bandwidth achieved by this method. This additional 200 Hz of bandwidth is vital for listeners to be able to distinguish between sounds concentrated in high frequencies (e.g., the affricate sounds differentiating "chew" and "jew").

The YOHO database is the only large scale, scientifically controlled and collected, high-quality speech database for speaker authentication testing at high confidence levels. Table II-2 describes the YOHO database (Higgins 1990).

TABLE II-2
THE YOHO DATABASE

“Combination lock” phrases (e.g., 36-24-36)
186 subjects: 150 males, 36 females
Collected over 3 month period in a real-world office environment
4 enrollment sessions per subject with 24 phrases per session
~10 test sessions per subject with 4 phrases per session
Total of 1900 validated test sessions
8 kHz sampling with 3.8 kHz analog bandwidth
1.5 gigabytes of data

In a text-dependent speaker verification scenario, phrases are prompted and the claimant is requested to say them. The syntax used in the YOHO database is “combination lock” phrases. For example, the prompt might read: “Say: thirty-six, twenty-four, thirty-six.” Where the claimant is to speak the phrase as three doublets.

The U.S. Government is very interested in improving speaker authentication performance with “clean” data. However, there is an enormous consumer market that deals with noisy corrupted data (e.g., telephone services).

Speech Production

There are two main sources of speaker-specific characteristics of speech: physical and learned. Vocal tract shape is an important physical distinguishing factor of speech. The vocal tract is generally considered as the speech production organs above the vocal folds. As shown in Figure II-1 (Flanagan 1972), this includes the laryngeal pharynx (beneath epiglottis), oral pharynx (behind tongue, between epiglottis and velum), oral cavity (forward of the velum and bounded by the lips, tongue, and palate), nasal pharynx (above velum, rear end of nasal cavity), and the nasal cavity (above the palate and extending from the pharynx to the nostrils). An adult male vocal tract is approximately 17 cm long (Flanagan 1972).

The vocal folds (also known as vocal cords) are shown in Figure II-1. The larynx is composed of the vocal folds, the top of the cricoid cartilage, the arytenoid cartilages, and the thyroid cartilage (also known as “Adam’s apple”). The vocal folds are stretched between the thyroid cartilage and the arytenoid cartilages. The area between the vocal folds is called the glottis.

As the acoustic wave passes through the vocal tract, its frequency content (spectrum) is altered by the resonances of the vocal tract. Vocal tract resonances are called *formants*. Thus, the vocal tract shape can be estimated from the spectral shape (e.g., formant location and spectral tilt) of the voice signal.

Voice verification systems typically use features derived only from the vocal tract. As seen in Figure II-1, the human vocal mechanism is driven by an excitation source which also contains speaker-dependent information. The excitation is generated by airflow from the lungs, carried by the trachea (also called the “wind pipe”), through the vocal folds (or the arytenoid cartilages). The excitation can be characterized as phonation, whispering, friction, compression, vibration, or a combination of these.

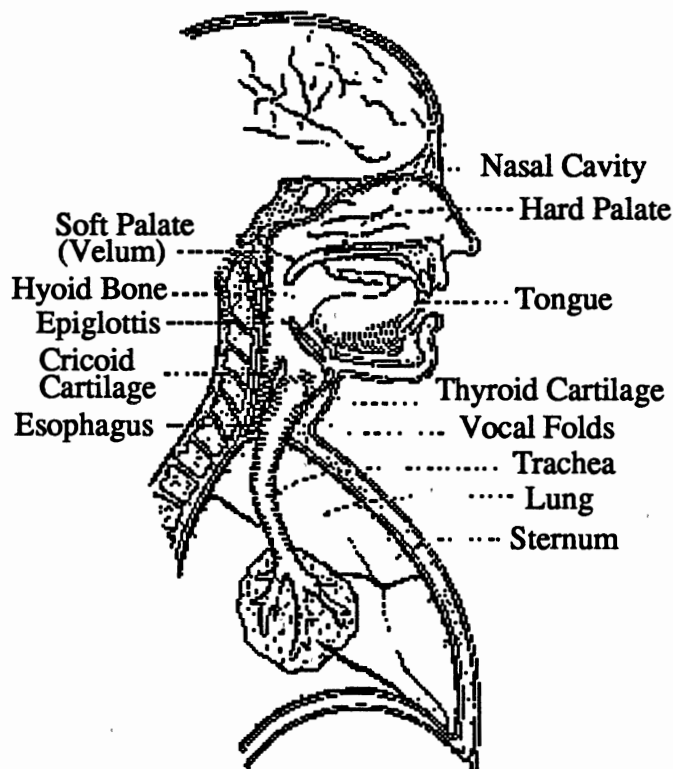


Figure II-1. Human Vocal System

During phonation, air flow is modulated by the vocal folds. When the vocal folds are closed, pressure builds up underneath them until they blow apart. Then, the folds are drawn back together again by their tension, elasticity, and the Bernoulli effect. The pulsed airstream, arising from the oscillating vocal folds, excites the vocal tract. The frequency of oscillation is called the fundamental frequency and it depends upon the length, tension, and mass of the vocal folds. Thus, fundamental frequency is another distinguishing characteristic which is physically based.

Whispered excitation is produced by airflow rushing through a small triangular opening between the arytenoid cartilages at the rear of the nearly closed vocal folds. This results in turbulent airflow, which has a wide-band noise characteristic (Parsons 1987).

Frication excitation is produced by constrictions in the vocal tract. The place, shape, and degree of constriction determines the shape of the broadband noise excitation. As the constriction moves forward, the spectral concentration generally increases in frequency. Sounds generated by frication are called *fricatives* or *sibilants*. Frication can occur without phonation (e.g., “s” as in sass) or with phonation (e.g., “z” as in zoos).

Compression excitation results from releasing a completely closed and pressurized vocal tract. This results in silence (during pressure accumulation) followed by a short noise burst. If the release is sudden, a *stop* or *plosive* is generated. If the release is gradual, an *affricate* is formed.

Vibration excitation is caused by air being forced through a closure other than the vocal folds, especially at the tongue (e.g., trilled “r”).

Speech produced by phonated excitation is called *voiced*; while other types of excitation produce *unvoiced* speech (phonation plus frication is called *mixed voiced*). Because of the differences in the manner of production, it’s reasonable to expect some speech models to be more accurate for certain classes of excitation than others. Unlike phonation and whispering, the points of frication, compression, and vibration excitations are actually inside the vocal tract, itself. This could cause difficulties for models that assume an excitation at the bottom end of the vocal tract. For example, the linear prediction model assumes a vocal tract excited at a closed end. Phonation excitation is the only one that approximates this assumption. Thus, it’s reasonable to use different models or different weighting for those regions of speech that violate the model assumptions.

The respiratory (thoracic area) plays a role in the resonance properties of the vocal system. The trachea is a pipe, typically 12 cm long and 2 cm in diameter, made up of rings of cartilage joined by connective tissue joining the lungs and the larynx. When the vocal folds are in vibration, there are resonances above and below the folds. Subglottal resonances are largely dependent upon the properties of the trachea (Pentz 1990).

Because of this physiological dependence, subglottal resonances have speaker-dependent properties. For this reason, subglottal resonances are pursued in this research.

Other physiological speaker-dependent properties include: vital capacity (the maximum volume of air you can blow out after maximum intake), maximum phonation time (the maximum duration a syllable can be sustained), the phonation quotient (ratio of vital capacity to maximum phonation time), glottal air flow (amount of air going through vocal folds). Because sound and airflow are different, these dimensions may be difficult to acquire from the acoustic signal alone.

Other aspects of speech production that could be useful for discriminating between speakers are learned characteristics, including speaking rate, prosodic effects, and dialect (which might be captured spectrally as a systematic shift in formant frequencies).

Linear Prediction

The all-pole linear predictor models a signal, s_n , by a linear combination of its past values and a scaled present input (Makhoul 1975):

$$s_n = -\sum_{k=1}^p a_k \cdot s_{n-k} + G \cdot u_n \quad (\text{II-1})$$

where s_n is the present output, p is the prediction order, a_k are the model parameters called the predictor coefficients (PCs), s_{n-k} are past outputs, G is a gain scaling factor, and u_n is the present input. In speech applications, the input, u_n , is generally unknown, so it's ignored. Therefore, the linear prediction approximation, \hat{s}_n , depending only on past output samples, is:

$$\hat{s}_n = -\sum_{k=1}^p a_k \cdot s_{n-k} \quad (\text{II-2})$$

This greatly simplifies the problem of estimating the a_k because the source (i.e., the glottal input) and filter (i.e., the vocal tract) have been decoupled. The source, u_n , which corresponds to the human vocal tract excitation is not modeled by these PCs. It is certainly reasonable to expect that some speaker-dependent characteristics are present in this excitation signal (e.g., fundamental frequency). Therefore, if the excitation signal is ignored, valuable speaker authentication discrimination information could be lost.

Defining the prediction error, e_n (also known as the residual), as the difference between the actual value, s_n , and the predicted value, \hat{s}_n , yields:

$$e_n = s_n - \hat{s}_n = s_n + \sum_{k=1}^p a_k \cdot s_{n-k} \quad (\text{II-3})$$

Therefore, the prediction error, e_n , is identical to the scaled input signal, $G \cdot u_n$. Letting E represent the mean squared error (MSE):

$$E = \sum_n e_n^2 = \sum_n \left[s_n + \sum_{k=1}^p a_k \cdot s_{n-k} \right]^2 \quad (\text{II-4})$$

The minimum MSE criteria resulting from:

$$\frac{\partial E}{\partial a_k} = 0, \quad \forall \quad i = 1, 2, \dots, p \quad (\text{II-5})$$

is:

$$\sum_{k=1}^p a_k \cdot \sum_n s_{n-k} s_{n-i} = - \sum_n s_n s_{n-i}, \quad \forall \quad i = 1, 2, \dots, p \quad (\text{II-6})$$

The summation ranges on n have been intentionally omitted for generality. If the summation is of infinite extent (or over the nonzero length of a finite extent window (Harris 1978)), the summations on s are the autocorrelations at lags $i - k$ for the left sum

and at lag i for the right sum. This results in the “autocorrelation method” of linear prediction (LP) analysis. (Other methods, such as “covariance” and Burg’s, arise from variations on windowing, the extent of the signal, and whether the summations on s are one or two sided.) The time-averaged estimates of the autocorrelation at lag τ can be expressed as:

$$R_{\tau} = \sum_{i=0}^{N-1-\tau} s(i) \cdot s(i + \tau) \quad (\text{II-7})$$

The autocorrelation method yields the system of equations named after Yule’s pioneering all-pole modeling in sunspot analysis and given by Equation II-8.

$$\begin{bmatrix} R_0 & R_1 & R_2 & \cdots & R_{p-1} \\ R_1 & R_0 & R_1 & \ddots & R_{p-2} \\ R_2 & R_1 & R_0 & \ddots & R_{p-3} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ R_{p-1} & R_{p-2} & R_{p-3} & \cdots & R_0 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \\ a_p \end{bmatrix} = - \begin{bmatrix} R_1 \\ R_2 \\ R_3 \\ \vdots \\ R_p \end{bmatrix} \quad (\text{II-8})$$

The LP model parameters we seek are the a_k . For a p^{th} order prediction, the speech signal is modeled by a p dimensional a_k vector and, as the Yule-Walker equations show, this requires the computation of $p + 1$ autocorrelations and matrix inversion. The matrix inversion problem is greatly simplified because of the symmetric Toeplitz autocorrelation matrix, $R_{i,j} = R_{|i-j|}$, and the form of the autocorrelation vector, which are exploited by Durbin’s recursive algorithm. This algorithm is the most efficient method known for solving this particular system of equations (Makhoul 1975): Note that in the process of solving for the predictor coefficients, a_k , of order p , the a_k for all orders less than p are obtained with their corresponding mean-square prediction error: $MSE_i = E_i / R_0$. In each recursion of Durbin’s algorithm, the prediction order is increased and the corresponding error is determined; this can be monitored as a stopping criteria on the prediction order, p . Durbin’s procedure is so efficient that it requires only roughly an eighth of the

$$\begin{aligned}
E_0 &= R_0 \\
k_i &= -\left[R_i + \sum_{j=1}^{i-1} a_j^{(i-1)} R_{i-j} \right] / E_{i-1} \quad \forall 1 \leq i \leq p \\
a_i^{(i)} &= k_i \\
a_j^{(i)} &= a_j^{(i-1)} + k_i a_{i-j}^{(i-1)} \quad \forall 1 \leq j \leq i-1 \\
E_i &= (1 - k_i^2) E_{i-1} \\
a_j &= a_j^{(p)} \quad \forall 1 \leq j \leq p
\end{aligned}
\left. \vphantom{\begin{aligned} E_0 &= R_0 \\ k_i &= -\left[R_i + \sum_{j=1}^{i-1} a_j^{(i-1)} R_{i-j} \right] / E_{i-1} \\ a_i^{(i)} &= k_i \\ a_j^{(i)} &= a_j^{(i-1)} + k_i a_{i-j}^{(i-1)} \\ E_i &= (1 - k_i^2) E_{i-1} \\ a_j &= a_j^{(p)} \end{aligned}} \right\} \quad \forall i = 1, 2, \dots, p \quad (\text{II-9})$$

operations required to compute the autocorrelations (Fussell 1986).

Using the a_k model parameters, the following equation represents the fundamental basis of LP representation. It implies *any* signal is defined by a linear predictor and the corresponding linear prediction error. Obviously, the residual contains all the information not contained in the PCs.

$$s_n = -\sum_{k=1}^p a_k \cdot s_{n-k} + e_n \quad (\text{II-10})$$

From Equation II-1, the LP transfer function is defined as:

$$H(z) \equiv \frac{S(z)}{U(z)} \equiv \frac{Z[s_n]}{Z[u_n]} \quad (\text{II-11})$$

which yields:

$$H(z) = \frac{G}{1 + \sum_{k=1}^p a_k z^{-k}} \equiv \frac{G}{A(z)} \quad (\text{II-12})$$

where $A(z)$ is known as the p^{th} order inverse filter.

LP analysis determines the PCs of the inverse filter, $A(z)$, that minimize the prediction error, e_n , in some sense. Typically, the MSE is minimized because it allows a simple, closed-form solution of the PCs. Minimizing MSE error tends to produce a flat

(band-limited white) magnitude spectrum of the error signal. Hence, the inverse filter, $A(z)$, is also known as a “whitening” filter. This band-limited whitened spectrum leads to a narrow (nearly impulsive) *sinc* pulse in the time domain.

If a voiced speech signal “fits the model,” then the residual consists of a pitch periodic impulse train. Therefore, the maximum prediction errors (residual peaks) occur at the pitch rate. (Many pitch detection algorithms exploit this property.) Thus, in the time domain, the majority of information lost in the PCs occurs in the vicinity of these “pitch peaks.”

Features are constructed from the speech model parameters; for example, the a_k , above. In this research, the linear prediction coefficients are estimated on unpreemphasized speech sampled at 8 kHz every 10 ms using a 10th order autocorrelation analysis method with 20 ms overlapping Hamming windows and 15 Hz bandwidth expansion. The bandwidth expansion operation replaces the LP analysis predictor coefficients, a_k , by $a_k \gamma^k$, where $\gamma = 0.994$ for a 15 Hz expansion. This broadens the formant bandwidths by shifting the poles radially toward the origin in the z -plane by the weighting factor, γ , for $0 < \gamma < 1$. These LP coefficients are typically nonlinearly transformed into perceptually meaningful domains suited to the application. Some domains useful for speech coding and recognition include: reflection coefficients (RCs); log-area ratios (LARs) or arcsin of the RCs; LP cepstrum (Rabiner and Schafer 1978); and line spectrum pair frequencies, recently introduced by Itakura (Itakura 1975; Saito and Nakata 1985).

Reflection Coefficients

If Durbin’s algorithm is used to solve the LP equations, the reflection coefficients are the intermediate k , variables in the recursion. The reflection coefficients can also be obtained from the LP coefficients using the backward recursion (Rabiner and Schafer 1978):

$$\left. \begin{aligned} \alpha_j^{(p)} &= a_j \\ k_i &= \alpha_i^{(i)} \\ \alpha_j^{(i-1)} &= \frac{\alpha_j^{(i)} + \alpha_i^{(i)} \cdot \alpha_{i-j}^{(i)}}{1 - k_i^2} \quad \forall \quad 1 \leq j \leq i-1 \end{aligned} \right\} \quad \forall \quad i = p, p-1, \dots, 1 \quad (\text{II-13})$$

Log Area Ratios

The vocal tract can be modeled as an electrical transmission line, a waveguide, or an analogous series of cylindrical acoustic tubes. At each junction, there can be an impedance mismatch or an analogous difference in cross-sectional areas between tubes. At each boundary, a portion of the wave is transmitted and the remainder is reflected (assuming lossless tubes). The reflection coefficients, k_i , are the percentage of the reflection at these discontinuities. If the acoustic tubes are of equal length, the time required for sound to propagate through each tube is equal (assuming planar wave propagation). Equal propagation times allow simple z-transformation for digital filter simulation. For example, a series of five acoustic tubes of equal lengths with cross-sectional areas A_0, A_1, \dots, A_5 could look like Figure II-2. This series of five tubes represents a fifth order system that might fit a vocal tract minus the nasal cavity. Given boundary conditions, the reflection coefficients are determined by the ratios of the adjacent cross-sectional areas (Rabiner and Schafer 1978). For an N^{th} order system, the boundary conditions given below correspond to a closed glottis (zero area) and a large area following the lips.

$$\begin{aligned} A_0 &= 0 \\ A_{N+1} &\gg A_N \\ r_k &= \frac{A_{k+1} - A_k}{A_{k+1} + A_k} \end{aligned} \quad (\text{II-14})$$

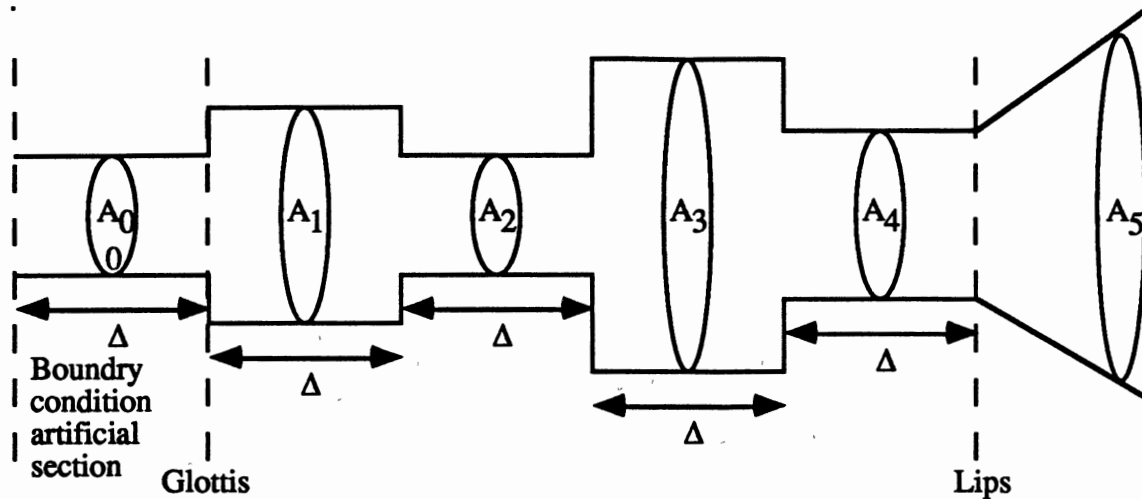


Figure II-2. Acoustic Tube Model of Speech Production

Thus, the reflection coefficients can be derived from an acoustic tube model or an autoregressive model. To paraphrase the late Professor Feynman, one measure of the degree of our understanding is the number of different ways in which we can arrive at the same result.

If the speech signal is preemphasized prior to LP analysis to compensate for the effects of radiation and the nonwhite glottal pulse, then the resulting cross-sectional areas are often similar to the human vocal tract configuration used to produce the speech under analysis (Rabiner and Schafer 1978). They cannot be guaranteed to match, however, because of the nonuniqueness properties of the vocal tract configuration. For example, to keep their lip opening small, ventriloquists exploit this property by compensating with the remainder of their vocal tract configuration.

Narrow bandwidth poles result in $|k_i| \approx 1$. Inaccurate representation of these RCs can cause gross spectral distortion. Taking the log of the area ratios results in more uniform spectral sensitivity. The LARs are defined as the log of the ratio of adjacent cross-sectional areas:

$$g_i = \log \left[\frac{A_{i+1}}{A_i} \right] = \log \left[\frac{1+k_i}{1-k_i} \right] = 2 \tanh^{-1} k_i \quad (\text{II-15})$$

Arcsin Reflection Coefficients

To avoid the singularity of the LARs at $k_i = 1$, while retaining approximately uniform spectral sensitivity, the arcsin of the RCs are a common choice:

$$g'_i = \sin^{-1} k_i \quad (\text{II-16})$$

Line Spectrum Pair Frequencies

The LSPs are a representation of the PCs of the inverse filter, $A(z)$, where the p zeros of $A(z)$ are mapped onto the unit circle in the z -plane through a pair auxiliary $p+1$ order polynomials: $P(z)$ (symmetric) and $Q(z)$ (antisymmetric) (Kang and Fransen 1985):

$$\begin{aligned} A(z) &= \frac{1}{2} [P(z) + Q(z)] \\ P(z) &= A(z) + z^{-(p+1)} A(z^{-1}) \\ Q(z) &= A(z) - z^{-(p+1)} A(z^{-1}) \end{aligned} \quad (\text{II-17})$$

where the LSPs are the frequencies of the zeros of $P(z)$ and $Q(z)$. By definition, a stable LP synthesis filter has all its poles inside the unit circle in the z -plane. The corresponding inverse filter is therefore minimum phase inverse because it has no poles or zeros outside the unit circle. Any minimum phase polynomial can be mapped by this transform to represent each of its roots by a pair of frequencies (phases) with unit magnitude. The LSP representation of the LP filter has a direct frequency domain interpretation that is especially useful in efficient (accurate and compact) coding and smoothing of the LP filter coefficients (Campbell and others 1991).

For example, an 8th order 8 kHz LP analysis of the vowel /U/ (as in foot) had the predictor coefficients shown in Table II-3.

TABLE II-3
EXAMPLE LINEAR PREDICTOR COEFFICIENTS

Power of z	0	-1	-2	-3	-4	-5	-6	-7	-8
Predictor Coefficient	1	-2.346	1.657	-0.006	0.323	-1.482	1.155	-0.190	-0.059

Evaluating the magnitude of the z-transform of $H(z)$ at equally spaced intervals on the unit circle yields the following power spectrum having formants (vocal tract resonances or spectral peaks) at 390 Hz, 870 Hz, and 3040 Hz (Figure II-3). These resonance frequencies are in agreement with the Peterson and Barney formant frequency data for the vowel /U/ (Rabiner and Schafer 1978).

Because the PCs are real, the Fundamental Theorem of Algebra guarantees that the roots of $A(z)$, $P(z)$, and $Q(z)$ will occur in complex conjugate pairs. Because of this conjugate property, the bottom half of the z-plane is redundant. The LSPs at 0 and π are always present by construction of P and Q . Therefore, the PCs can be represented by the number of LSPs equal to the prediction order, p , and are represented by the frequencies of the zeros of P and Q in the top-half z-plane (Figure II-4).

The LSPs satisfy an interlacing property of the zeros of the P and Q polynomials, which holds for all minimum phase $A(z)$ polynomials (Kang and Fransen 1985):

$$0 = \omega_0^{(Q)} < \omega_1^{(P)} < \omega_2^{(Q)} < \dots < \omega_{p-1}^{(P)} < \omega_p^{(Q)} < \omega_{p+1}^{(P)} = \pi \quad (\text{II-18})$$

Each complex zero of $A(z)$ maps into one zero in each $P(z)$ and $Q(z)$. When the $P(z)$ and $Q(z)$ frequencies are close, it is likely that the original $A(z)$ zero was close to the unit circle and a formant is likely to be in between the corresponding LSPs. Distant P and Q zeros are likely to correspond to wide bandwidth zeros of $A(z)$ and most likely contribute only to shaping or spectral tilt. Figures II-3 and II-4 demonstrate this behavior.

Speech Perception

The human hearing system's extraordinary dynamic range, speaker identification and speaker-independent speech recognition and understanding is nothing short of miraculous. The ear, shown in Figure II-5 (Slaney 1988) is a complex structure

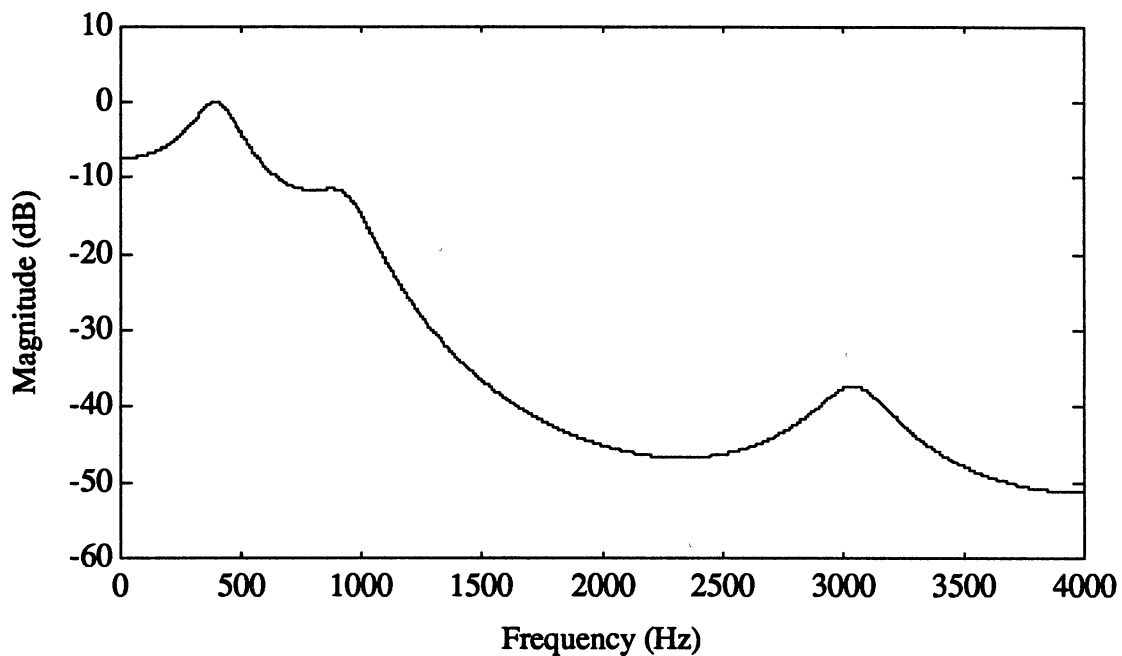


Figure II-3. Frequency Response

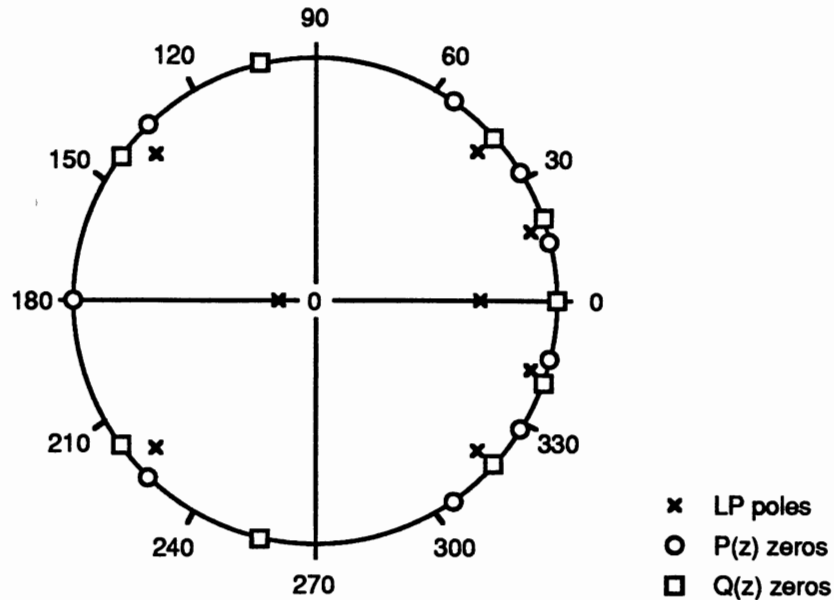


Figure II-4. LSP Frequencies and LP Poles in the z-Plane

The *outer ear* is a large directional acoustic horn. The ear canal leading from the outer to middle ear is open at the outer ear and closed at the ear drum and forms a quarter-wavelength Helmholtz resonator with its first resonance at approximately 3 kHz. The *middle ear* is the ear drum and the *ossicles* (three bones) which nonlinearly transmit sound to the inner ear's *oval window*. The oval window couples to the *cochlea*, which has a dividing *basilar membrane* forming two concentric "snail shells" (at the tip of the cochlea arrow in Figure II-5). The *organ of Corti* lies along the basilar membrane and has about 20,000 sensory hair cells. The endings of the *auditory nerve* terminate on these hair cells, each having about 100 hairs that bend from vibrations to cause neural firings. The neural information then ascends to the brain (Fussell 1986).

The basilar membrane varies in shape and tautness along its length. Vibrations at different frequencies excite different regions of hair cells. Regions of hair cells responding to a particular frequency of vibration are labeled by this characteristic

frequency. The organ of Corti produces electrical potentials, called the cochlear microphonic, which represent the acoustic signal (Martin 1991).

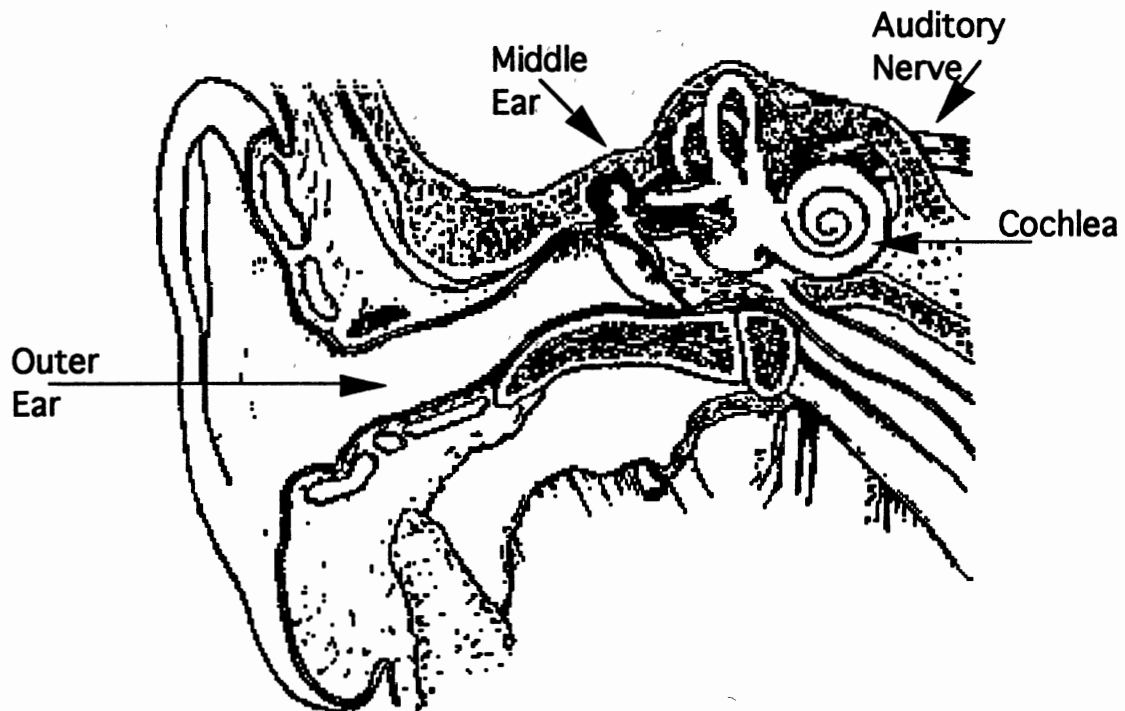


Figure II-5. The Human Ear

Except in the remotest sense, it would be naive to think that the simple features discussed so far could capture the subtleties and complexities of the human hearing system. Therefore, the next section will explicitly set out to find features that are more directly related to the human hearing system.

Perceptually Motivated Features

Although LP-derived features are common, other popular feature sets have a greater perceptual influence, such as critical-band filter-bank parameters and mel-warped cepstrum (the Fourier transform of a log magnitude spectrum which has been frequency warped according to the mel scale). One would expect that strong feature selection methods combined with speech perception and speech production based features would yield powerful new speaker discriminatory feature sets. However, one must be cautioned that attempting to imitate the human verification system is not necessarily the optimum solution. For example, airplanes don't flap their wings!

To make speech comparisons in a perceptually meaningful domain and greatly reduce storage and computation, speaker authentication systems extract features based on parametric models of the speech signal, rather than using the raw digital signal, itself. Features are extracted from the speech signal to construct speaker models during enrollment and for use in comparison with those models during authentication. The most common speech signal parameterization begins with LP analysis to derive a vector of PCs with much lower dimension than the input (typically, the input is \mathcal{R}^{100} and the PCs are \mathcal{R}^{10}). The PC vector is then nonlinearly transformed to form the feature vector in a domain where simple distance measures relate to the application (e.g., inter- versus intraspeaker variability). The search for the ultimate feature vector remains elusive and, as can be seen from the literature, there is no universally agreed upon "best" speech feature vector. It's especially ironic that cepstrum based features are commonly used for both speaker recognition and speech recognition because what's considered information for one is noise for the other and vice versa.

Table II-4 lists some commonly used speech processing features. As described below, not all of these features are desirable for speaker verification. Fundamental frequency might come to mind when we think of someone's "pitch" as helpful speaker identifying

TABLE II-4
COMMON SPEECH PROCESSING FEATURES

Fundamental frequency
Short term energy
Zero crossing rate
Short term spectrum:
• LP coefficients
• Nonlinear transforms of LP coefficients (e.g., reflection coefficients)
• Cepstrum, possibly mel warped

information. Unfortunately, it exhibits large intraspeaker variability and is strongly influenced by the subject's mood; thus, it cannot be used by itself as a reliable feature. The short term energy represents the dynamics of a person's speech, but is also somewhat mood dependent and is also too weak to use by itself. The zero crossing rate represents the dominant spectral component and could be useful. Measures related to the short term spectrum will be shown later to correlate with the speaker's vocal tract configuration. Because the speech signal is stationary only on a short-time basis, the short-term spectral characteristics are a powerful measure. This is much less sensitive to the speaker's mood and is the basis of most feature sets in use today for speaker authentication.

Feature vectors can be constructed by concatenations of these features. For example, one might consider a feature vector consisting of the short term energy and 10 reflection coefficients. The features could also be speaker dependent or adaptive (Attili and others 1988).

Pitch

Acoustic dimensions such as intensity and frequency can be measured. The perceptual correlates to these dimensions are loudness and pitch, which are determined by subjective psychoacoustic experiments. In an attempt to model human hearing, various measurement approximations have been tried to mimic perception, as shown in Table II-5.

TABLE II-5
ACOUSTIC AND PERCEPTUAL CORRELATES

Acoustic		Perceptual	
Dimension	Units	Dimension	Units
Intensity	dB SPL $\left(2 \cdot 10^{-4} \frac{\text{dyne}}{\text{cm}^2}\right)$	Loudness	Phon (equal) Sone (scaling)
Frequency	Hz	Pitch	mel (scaling)

The mel scale is the result of a psychoacoustic experiment where subjects are asked to judge if one tone is “half as high in pitch as another.” The resulting mel scale frequency warping nonlinearly maps the input frequency in Hertz to subjective pitch in mels. This scale is referenced to 1 kHz = 1000 mels. As shown in Figure II-6, for equal intervals on the mel axis, the listener perceives equal pitch ratios projected through the curve to the Hz axis (Borden and Harris 1984) and is closely approximated (Neuburg 1981) by:

$$mels \approx 1000 \cdot \log_2 \frac{1+f}{1000} \quad (\text{II-19})$$

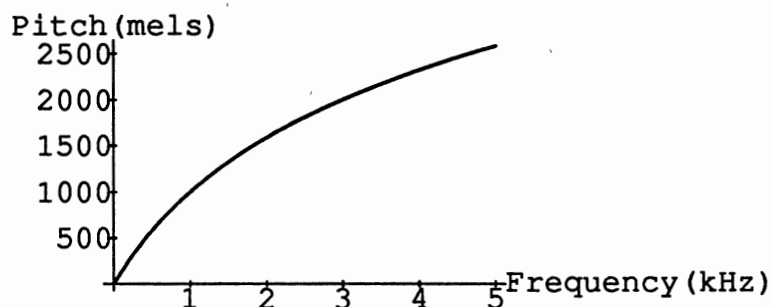


Figure II-6. The mel Pitch Scale

The Bark scale is the result of a psychoacoustic experiment where a pure tone is masked (i.e., inaudible) by a band of noise centered in frequency on the tone. As the bandwidth of the noise increases, the amplitude of the noise needed to just mask the tone decreases up to the critical bandwidth. Beyond this critical bandwidth (which depends on the tone frequency), the noise amplitude needed to mask the tone remains constant and independent of the noise bandwidth. Figure II-7 shows an approximation of the Bark scale (Neuburg 1981) by:

$$Barks \approx 7 \sinh^{-1} \frac{f}{650} \quad (\text{II-20})$$

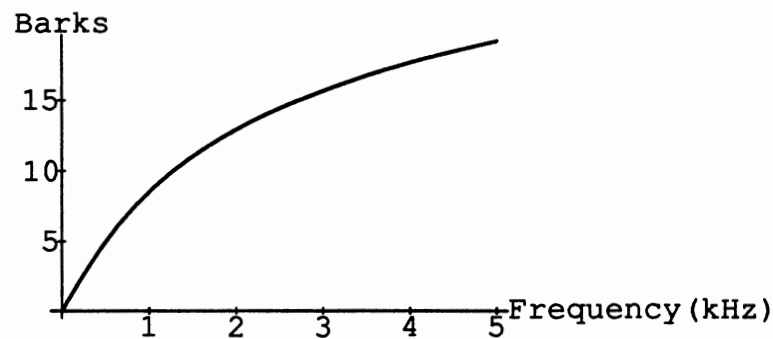


Figure II-7. The Bark Masking Scale

Its shape bears very close resemblance to the mel scale. In applications where only the relative ratio of mels or Barks is of importance, these two frequency scales can be considered identical (Neuburg 1981).

Auditory Model Pitch Estimation

In this research, voicing and pitch frequency are estimated by a new state-of-the-art auditory model-based pitch extractor, called AMPEX (Van Immerseel and Martens 1992). As shown in Figure II-8 (Hermes 1992), AMPEX performs a temporal analysis using delayed decisions (e.g., dynamic program) of the outputs emerging from a new auditory model. The auditory components modeled are the outer and middle ear chain, filtering in the cochlea, mechanical-to-neural transduction (with short-time adaptation in the hair cells), and auditory nerve transmission. AMPEX was found to outperform other methods, including the subharmonic summation method without delayed decisions (Hermes 1988; Hermes 1992); however, its computational burden is immense.

After tuning the parameters of the AMPEX and SHS algorithms for optimum performance on the YOHO database, AMPEX's pitch track was found to be closer to the

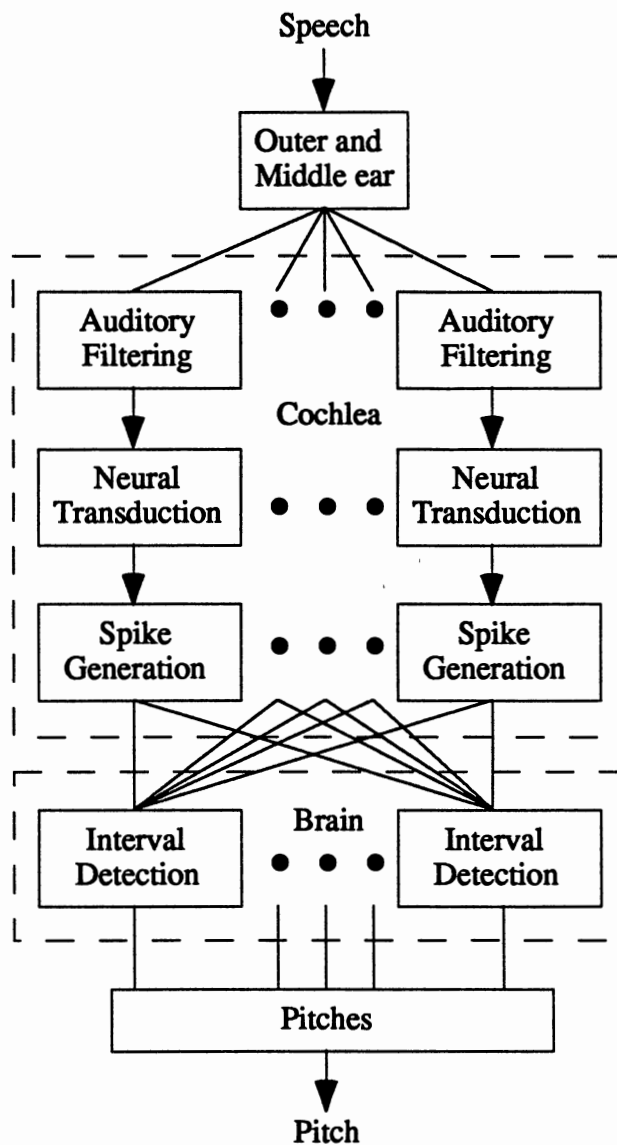


Figure II-8. Auditory Model Pitch Extractor

subjective true pitch for speech from the database. The band-pass filtering (BPF) of the YOHO database appears to cause errors in the pitch estimated by SHS. The lack of a delayed decision (e.g., dynamic program) caused additional errors in the SHS pitch track. The parameters optimized in SHS were adjustments to the high frequency deweighting factor, the weighting function, the down sampling filter, and FFT length. The parameters

optimized in AMPEX were the input level scaling factor and minimum evidence of voicing threshold. The input level sensitivity of AMPEX is a potential weakness in practical implementations, but, fortunately, its setting doesn't appear to be too critical.

To try many feature extraction variations on the YOHO database (about 20 hours of speech), AMPEX had to process the speech data files quickly (preferably faster than real time). With full optimization, Sun's C compiler (cc) generated code that was 9.7 times real time on a Sun SPARC-2 workstation computer. Sun's ANSI C compiler (acc) and the GNU C compiler (gcc) did a little better at 8 times real SPARC-2 time, but this was still too slow. Using one head on a Cray-2, AMPEX ran in 4.4x real time (the AMPEX C code vectorized poorly). Although the Cray-2 ran AMPEX faster than a single SPARC2, it was still too slow. A throughput of 0.2x real time was achieved using a suite of Bourne-shell scripts to multiprocess AMPEX across a network of 40 Sun SPARC-2s! AMPEX is well suited to multiprocessing because it consumes very little memory (few page faults) and was able to run at low priority on people's SPARC-2s without them noticing.

Loudness

Perceived loudness is at least a function of both frequency and level. A phon level is the result of a psychoacoustic experiment where listeners are asked to adjust the amplitude of various tones to match the amplitude of a 1 kHz reference tone at a given amplitude. By conducting a series of experiments using different amplitude reference tones, a family of curves with contours of equal subjective loudness is found. These curves, shown in Figure II-9 (Borden and Harris 1984) are known as the Fletcher-Munson curves.

Equal-loudness compensation is an approximation of the ear's unequal sensitivity at different frequencies; i.e., frequency equalization. These curves show that maximum acuity (frequencies we are most sensitive to) occurs between 2 and 4 kHz. It's no

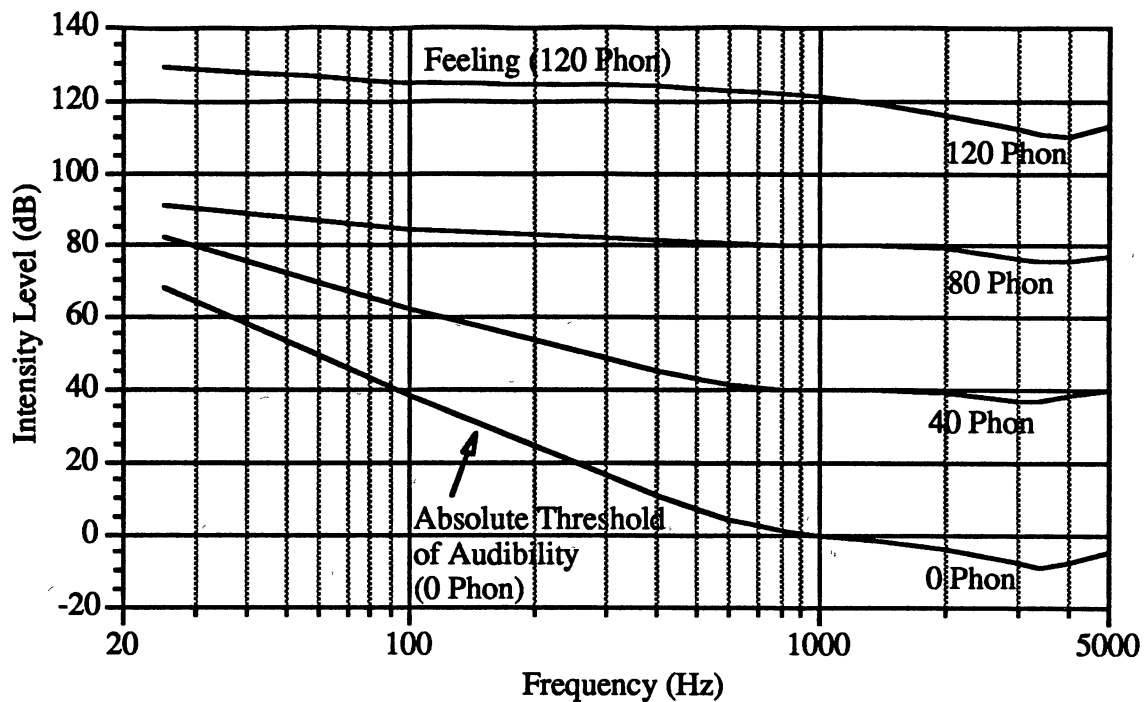


Figure II-9. Equal-Loudness Level Contours

coincidence that this is in the neighborhood of the Helmholtz resonance of the ear canal and that human speech evolved in this vicinity.

The loudness level scale, measured in sones, is the result of a psychoacoustic experiment where listeners are asked to set the loudness of a sound to $\frac{1}{2}$, 2, $\frac{1}{10}$, or 10 times the loudness of a 1 kHz reference tone. As shown in Figure II-10, the perceived loudness in sones is often approximated from the loudness level in phons (Parsons 1987) by:

$$L_s \approx 0.063 \cdot 10^{0.03L_p} \quad (\text{II-21})$$

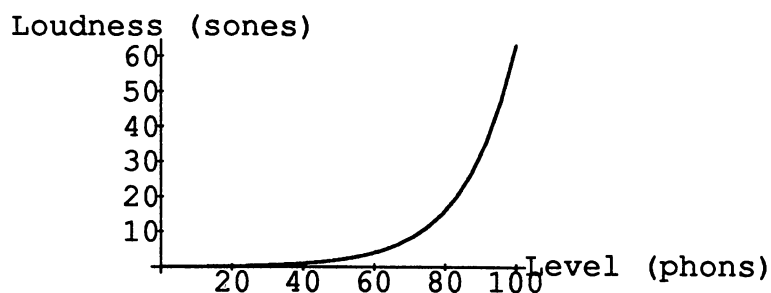


Figure II-10. Perceived Loudness Scale

The intensity-loudness power law is roughly a cube-root amplitude compression to approximate the power law of hearing by simulating the nonlinear relation between the intensity of sound and its perceived loudness (Hermansky 1990).

Perceptual-Model Filterbank

Combining all of the above, we can approximate the ear's frequency-dependent and amplitude-dependent responses to simple sounds such as pure tones. Complex sounds, such as speech, may need yet another level of understanding to accurately reflect their perceptual properties. The perceptual-model filterbank used in this research was adopted for an 8 kHz sampling frequency from the front-end proposed by Hermansky in his method of perceptual linear prediction (Hermansky 1990). The processing steps for this filterbank are described in Table II-6. In this research, the short-term power spectrum is estimated via the periodogram method (Oppenheim and Schaffer 1989) every 10 ms using 20 ms overlapping Hamming windows with 256 point FFTs. A fixed equal-loudness curve was selected to approximate the 40 phon level for the preemphasis. A cube-root nonlinearity is used for the compression to simulate the intensity-loudness power law of hearing. To sample the 0 to 4 kHz (0 to 15.6 Bark) analysis bandwidth at approximately

TABLE II-6
PERCEPTUAL-MODEL FILTERBANK

-
1. Estimate the short-term power spectrum.

 2. Convolve the power spectrum with a simulated critical-band masking pattern.

 3. Resample the critical-band power spectrum's frequency scale at approximately 1 Bark intervals via Hertz-to-Bark frequency warping to obtain a critical-band Bark-frequency power spectrum.

 4. Preemphasize the critical-band Bark-frequency power spectrum by a simulated equal-loudness curve to obtain a critical-band, Bark-frequency power spectrum in phons.

 5. Compress the critical-band, Bark-frequency power spectrum in phons through a simulated intensity-loudness power law to generate a perceptual-model filterbank feature vector in sones.
-

1 Bark intervals, 15 bands were chosen with 0.97344 Bark spacing. The magnitude response of the resulting perceptual-model filterbank (a 15 channel critical-band, Bark-frequency power spectrum in sones) is shown in Figure II-11. This perceptual-model filterbank accounts for the human ear's nonlinear transformations of frequency and amplitude and its analysis and masking behavior in response to complex sounds. Therefore, measures between perceptual-model filterbank feature vectors could correlate well with their perceptual closeness.

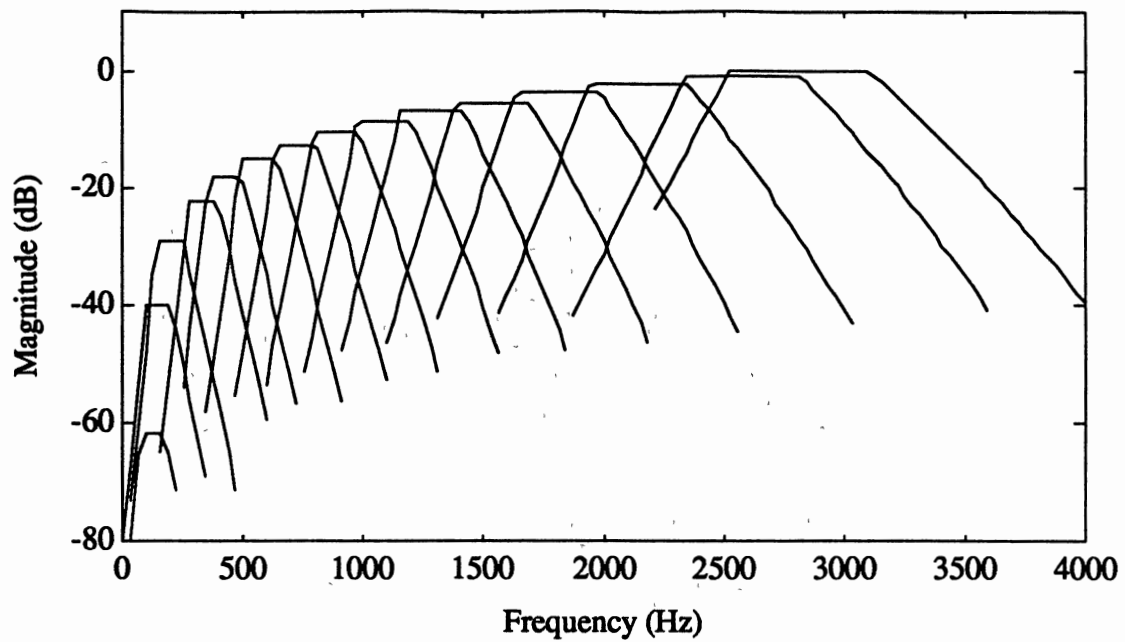


Figure II-11. Perceptual-Model Filterbank

The next chapter presents singular value decomposition in the context of linear prediction and speech perception.

CHAPTER III

LP AND SINGULAR VALUE DECOMPOSITION

Changes in the LP filter excitation on the synthetic speech output can be directly observed via singular value decomposition (SVD). A change in one of the right singular vectors (eigenvectors of $\mathbf{H}^T\mathbf{H}$) at the input produces a change in only one left singular vector (SV) at the output. The excitation waveform is expressed as a linear combination of the eigenvectors of the autocorrelation matrix ($\mathbf{H}^T\mathbf{H}$) of the LP filter's impulse response (\mathbf{h}). The LP filter convolution is transformed to a simple multiplication, providing important advantages in interpreting the role of each SV component in the excitation. This can be exploited in speech coding, fast vector quantization, code book design, and pattern matching.

Modern signal processing accounts for real-world observations that are incomplete and noisy. Traditional analysis techniques that assume stationarity and time invariance neglect these real-world conditions. SVD of the observation matrix allows robust separation of signal and noise spaces. SVD has given the modern counterparts of the traditional methods shown in Table III-1. Examples of SVD applied to system identification, signal detection, harmonic retrieval, principal component analysis, model reduction, detection of multiple sinusoids in noise, and its implementation can be found in (Deprettere 1988; Vaccaro 1991).

TABLE III-1
TRADITIONAL VERSUS MODERN METHODS

Traditional	Modern
Least Squares	Total Least Squares
L_2 Norm	Hankel-Norm Approximation
Fourier Transform	Prony-Type Modeling

SVD is generally the preferred method of rank determination. Many linear algebra theorems hinge upon a matrix being of full rank. Often the determination of rank is neglected. In the real world, rank determination is nontrivial because of noisy observations and numerical difficulties. SVD can be used to quantify how close a matrix is to rank deficiency (Golub and Van Loan 1983). For singular or nearly singular matrices, the best choice for solution of linear algebraic equations is almost always singular value decomposition with back substitution (Vetterling and others 1989). SVD is an excellent tool for linear fitting. The purpose of linear fitting is to reduce the data to a few model parameters, where there is linear dependence of the function on its fitting parameters (as opposed to its argument). SVD is generally the recommended method for linear fitting of a model function to a set of data because it never fails in practice, even in cases where near degeneracy of some basis functions occurs (which wreaks havoc with traditional linear least-squares methods) (Vetterling and others 1989). Now that we know a little about SVD's powers, let's review how it is found.

Definition of SVD

Singular value decomposition and eigenvalue-eigenvector decomposition (EVD) are intimately related. While EVD is restricted to square matrices, SVD can be used on both square and nonsquare matrices. By definition, the EVD of a symmetric matrix, $A = QAQ^T$, yields eigenvalues in the diagonal matrix Λ and an orthogonal eigenvector matrix Q (Strang 1988). The EVD doesn't exist for rectangular matrices, but if the left and right matrices are allowed to be any two orthogonal matrices, as opposed to transposes of each other, a decomposition can be performed. Furthermore, the diagonal matrix (now rectangular) can be made nonnegative (Strang 1988). This is the spirit of SVD.

Although SVD can be performed in the complex field, we'll restrict our results to the real field. (With minor alteration, these results can be extended to the complex field.) The literature abounds with proofs of these results (e.g., (Golub and Van Loan 1983), (Strang 1988)), so they won't be repeated here. For any real $m \times n$ matrix, A , of rank r :

$$A \in \mathcal{R}^{m \times n} \quad \text{rank}(A) = r \quad (\text{III-1})$$

there exist orthogonal matrices, U and V , and a diagonal ($a_{ij} = 0 \forall i \neq j$), strictly positive matrix, Σ :

$$U \in \mathcal{R}^{m \times m} \quad \Sigma \in \mathcal{R}^{m \times n} \quad V \in \mathcal{R}^{n \times n} \quad (\text{III-2})$$

such that A can be decomposed as:

$$A = U\Sigma V^T \quad (\text{III-3})$$

This is the *singular value decomposition* of \mathbf{A} . (It should be noted that if $m \geq n$, then an “economical form” exists, where \mathbf{U} is the same shape as \mathbf{A} and $\mathbf{\Sigma}$ is square (Moler and others 1989)).

$$\begin{aligned}
 \mathbf{U} &= [\mathbf{u}_1, \dots, \mathbf{u}_m] & \mathbf{u}_i &\in \mathcal{R}^m \\
 \mathbf{V} &= [\mathbf{v}_1, \dots, \mathbf{v}_n] & \mathbf{v}_i &\in \mathcal{R}^n \\
 \mathbf{\Sigma} &= \begin{bmatrix} \mathbf{S} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \in \mathcal{R}^{m \times n} & \mathbf{S} &= \text{diag}(\sigma_1, \dots, \sigma_r) \\
 & & & \sigma_1 \geq \dots \geq \sigma_r > 0
 \end{aligned} \tag{III-4}$$

The i^{th} diagonal element, σ_i , of the matrix, \mathbf{S} , is called the i^{th} *singular value* of \mathbf{A} . The number of singular values is equal to r , the rank of matrix, \mathbf{A} . The columns of \mathbf{U} are called the *left singular vectors* of \mathbf{A} . The columns of \mathbf{V} are called the *right singular vectors* of \mathbf{A} .

For symmetric positive definite matrices, the SVD degenerates to EVD, $\mathbf{A} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^T$. For indefinite matrices, any negative eigenvalues in \mathbf{A} become positive singular values in $\mathbf{\Sigma}$, which forces the left and right SVD matrices, \mathbf{U} and \mathbf{V} , to be different. As shown in Table III-2, the columns of \mathbf{U} and \mathbf{V} give orthonormal bases for all four fundamental subspaces (Strang 1988).

What does the SVD do geometrically? Given a sphere in n -dimensional space, if each vector in it is multiplied by an $m \times n$ matrix, \mathbf{A} , an ellipsoid in m -dimensional space results. The singular values of \mathbf{A} are the lengths of the principal axes of the ellipsoid and the left singular vectors of \mathbf{A} are the directions of the principal axes. If the matrix, \mathbf{A} , is singular, this will be reflected in the shape of the ellipsoid (Wolfram 1988).

The SVD can always be done, no matter how singular the matrix is, and it is “almost unique.” If a particular singular value is distinct, then the corresponding left and right singular vectors are also unique (Endsley 1991). Furthermore, the SVD is unique up to (1) making the same permutation of the columns of \mathbf{U} , elements of $\mathbf{\Sigma}$, and columns of \mathbf{V}

TABLE III-2
FUNDAMENTAL SUBSPACES

Columns of U or V	SVD Matrix	Subspace of A
first r	U	column space
last $m - r$	U	left null space
first r	V	row space
last $n - r$	V	null space

(or rows of V^T) or (2) forming linear combinations of any columns of U and V whose corresponding elements of Σ happen to be exactly equal (Press and others 1990).

So, how can U, V, and Σ be determined and computed? First, let's determine Σ :

$$\begin{aligned}
 A^T A &= (U \Sigma V^T)^T U \Sigma V^T \\
 &= V \Sigma^T U^T U \Sigma V^T & U^T U &= I \\
 &= V (\Sigma^T \Sigma) V^T & V V^T &= I \rightarrow V^T = V^{-1} \\
 &= V \Sigma^T \Sigma V^{-1} & V^{-1}(\bullet) &= V \\
 \Sigma^T \Sigma &= V^{-1} A^T A V
 \end{aligned} \tag{III-5}$$

$A^T A = V(\Sigma^T \Sigma)V^T$ is in the form of an EVD, so the modal matrix of $A^T A$ is given by V.

Thus, the columns of V, the right singular vectors of A, are the orthonormal eigenvectors of $A^T A$. Likewise, it's easy to show that the modal matrix of $AA^T = U \Sigma \Sigma^T U^T$ is given by U (Hershey and Yarlagaadda 1986). Thus, the columns of U, the left singular vectors of A, are the orthonormal eigenvectors of AA^T . It should be noted that EVD solutions have arbitrary scaling, so, if we restrict ourselves to

orthonormal solutions, there is a sign ambiguity between an eigenvalue and its corresponding eigenvector.

$\Sigma^T \Sigma = V^{-1} A^T A V$ is a similarity transform; thus, the matrices $\Sigma^T \Sigma$ and $A^T A$ are said to be *similar*. Similar matrices have the same eigenvalues (Strang 1988). Since Σ is diagonal, the matrix $\Sigma^T \Sigma$ is also diagonal and contains the elements $diag(\sigma_1^2, \dots, \sigma_r^2)$. Thus, the singular values are equal to the positive square roots of the eigenvalues of $A^T A$ (or AA^T , which has the same nonzero eigenvalues (Hershey and Yarlagadda 1986)).

Note, this would be a poor numerical method to actually calculate singular values because of the precision lost by the matrix squaring operation. Householder reduction to bidiagonal form and diagonalization by QR procedure with shifts is a traditional SVD serial-computation method (Press and others 1990). A new method, yielding very accurate singular values, has been proposed for inclusion in LINPACK (Demmel and Kahan 1990). For parallel computation, Jacobi methods (one or two sided) are usually preferred (Deprettere 1988).

Now, notice that A can be expanded into a sum of outer products using the SVD:

$$A = U \Sigma V^T = \sum_{i=1}^r \mathbf{u}_i \sigma_i \mathbf{v}_i^T = \sum_{i=1}^r \sigma_i \mathbf{u}_i \mathbf{v}_i^T \quad (\text{III-6})$$

Thus, any rectangular matrix can be represented as a weighted sum of r rank one matrices. Because the singular vectors \mathbf{u}_i and \mathbf{v}_i are normalized, the rank one matrices, $\mathbf{u}_i \mathbf{v}_i^T$, have the same Frobenius norm. So, the relative importance of each of the rank one matrices is determined by its corresponding singular value. This important result is used in reduced rank matrix approximations and will be used later.

As with vectors, norms are used to quantify the size of a matrix. The SVD is intimately related to various matrix norms. The squared 2-norm and squared Frobenius-norm of A can be expressed via SVD:

$$\begin{aligned}\|A\|_2^2 &\equiv \text{maximum eigenvalue of } A^T A = \sigma_1^2 \\ \|A\|_F^2 &\equiv \sum_{i,j} |a_{ij}|^2 = \sum_{i=1}^r \sigma_i\end{aligned}\quad (\text{III-7})$$

Note, some authors distinguish between the Frobenius, Euclidean, and Schur norms, while others do not (Deprettere 1988).

Only square, nonsingular matrices have inverses. Using singular value decomposition, however, it is possible to define a pseudoinverse even for nonsquare matrices or for singular square ones. The pseudoinverse of A , denoted $A^{(-1)}$, is often defined in terms of the SVD:

$$A^{(-1)} = V \Sigma^{(-1)} U^T \quad \Sigma^{(-1)} = \begin{bmatrix} S^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \in \mathcal{R}^{n \times m} \quad S^{-1} = \text{diag}\left(\frac{1}{\sigma_1}, \dots, \frac{1}{\sigma_r}\right) \quad (\text{III-8})$$

This definition of the pseudoinverse is sometimes known as the generalized inverse or the Moore-Penrose inverse. This pseudoinverse has the desirable property of minimizing the sum of the squares of the elements of $AA^{(-1)} - I$ or, equivalently,:

$$\|AA^{(-1)} - I\|_F^2 \quad (\text{III-9})$$

is minimized. The pseudoinverse found in this way is useful in performing fits to numerical data.

The *condition number* of A is given by the ratio of the largest to smallest singular values:

$$\kappa(A) \equiv \|A\| \|A^{-1}\| = \frac{\sigma_1}{\sigma_r} \geq 1 \quad (\text{III-10})$$

If $\kappa(A)$ is large, then A is said to be *ill conditioned*. If $\kappa(A)$ is small, then A is said to be *well conditioned*, which is desirable. Perfect conditioning, $\kappa(A) = 1$, implies that A is

orthogonal. The condition number measures the sensitivity of the solution of a system of linear equations to errors in the data and gives an indication of the accuracy of the results obtained from matrix inversion and solution of the set of linear equations. Because of the matrix squaring operation in EVD, SVD has smaller condition numbers. Thus, SVD is often preferable to an equivalent EVD (Deprettere 1988). Now that we've reviewed the SVD, let's apply it to an autoregressive (AR) model of speech synthesis.

SVD-Based Speech Models

Linear prediction is an autoregressive model. It is an extremely popular model of the short-term spectral envelope for speech signals. The spectral resonances (formants) of speech can be accurately represented by this all-pole model.

As shown in Figure III-1, the LP filter excited by x synthesizes a speech signal, s .

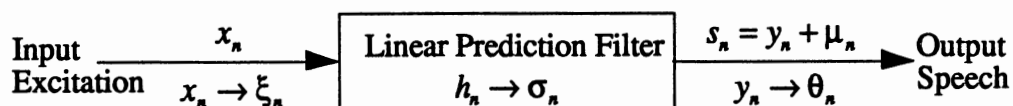


Figure III-1. LP Synthesis

Since the LP filter is AR, it has an infinite extent impulse response, h . At the n^{th} sampling instant, let x_n , h_n , and s_n represent the input, impulse response, and output, respectively. While the LP filter is time invariant (fixed h), the output can be represented by the infinite extent convolution:

$$s_n = \sum_{k=1}^{\infty} h_{n-k} x_k \quad n = 1, 2, \dots \quad (\text{III-11})$$

Notice that a change in any one sample of the input, x , is propagated into every sample of the output, s . Thus, a time domain analysis of how the input affects the output is very complicated. Because we'll eventually want a time varying LP filter, frequency domain analysis doesn't help, either. Let's see if SVD can help, but first we need to develop an LP matrix representation.

Matrix Form of Linear Prediction

Consider a frame of N speech samples over which the LP filter is time invariant. The convolution sum in Equation III-11 can be split in two, where one component of the response is due to excitation in the present and the other is due to excitation in the past:

$$\begin{aligned} s_n &= \sum_{k=1}^n h_{n-k} x_k + \sum_{k=n+1}^{\infty} h_{n-k} x_k & n = 1, 2, \dots, N \\ &= y_n + \mu_n & n = 1, 2, \dots, N \end{aligned} \quad (\text{III-12})$$

where y is the zero-state response (ZSR) and μ is the zero-input response (ZIR). In vector form:

$$\mathbf{s} = \mathbf{y} + \boldsymbol{\mu} \quad (\text{III-13})$$

where \mathbf{y} is the ZSR vector and $\boldsymbol{\mu}$ is the ZIR vector. Now, let's define an impulse response matrix:

$$\mathbf{H} \equiv \begin{bmatrix} h_0 & 0 & \cdots & 0 \\ h_1 & h_0 & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ h_{N-1} & h_{N-2} & \cdots & h_0 \end{bmatrix} \quad (\text{III-14})$$

then we can represent \mathbf{y} :

$$\begin{aligned} \mathbf{y} &= \sum_{k=1}^n h_{n-k} x_k & n = 1, 2, \dots, N \\ &= \mathbf{H}\mathbf{x} \end{aligned} \quad (\text{III-15})$$

Linear Prediction Properties

The lower-triangular Toeplitz matrix, \mathbf{H} , has some very interesting properties. For example, the determinant of \mathbf{H} is equal to h_0^N and, since $h_0 = 1$, $|\mathbf{H}| = 1$ (Atal 1989). This has interesting implications because the determinant of a matrix is equal to the product of all its eigenvalues.

$$|\mathbf{H}^T \mathbf{H}| = |\mathbf{H}^T| \cdot |\mathbf{H}| = 1 \Rightarrow \prod_j \lambda_j = 1 \quad (\text{III-16})$$

Recall that the singular values are equal to the positive square roots of the eigenvalues of $\mathbf{H}^T \mathbf{H}$:

$$\prod_j \lambda_j = 1 \Rightarrow \prod_i \sigma_i^2 = 1 \Rightarrow \prod_i \sigma_i = 1 \quad (\text{III-17})$$

Because the singular values are in descending order and their product equals one, the first singular value must be greater than or equal to 1:

$$\left. \begin{array}{l} \prod_i \sigma_i = 1 \\ \sigma_1 \geq \dots \geq \sigma_r > 0 \end{array} \right\} \Rightarrow \sigma_1 \geq 1 \quad (\text{III-18})$$

Therefore, a large singular value will force subsequent singular values to be small.

SVD of LP Impulse Response Matrix

Now that we have a matrix form, we're finally ready to apply the SVD to the square impulse response matrix, \mathbf{H} :

$$\begin{aligned} \mathbf{H} &= \mathbf{U}\Sigma\mathbf{V}^T \\ &= \sum_{i=1}^N \sigma_i \mathbf{u}_i \mathbf{v}_i^T \end{aligned} \quad (\text{III-19})$$

This follows from the important SVD result shown earlier, the $N \times N$ matrix, \mathbf{H} , can be expressed as a weighted sum of N rank-one matrices. It should be noted that if the left, \mathbf{u}_i , and right, \mathbf{v}_i , singular vectors are both multiplied by -1 , the summation still holds. Moreover, because of the Toeplitz structure of \mathbf{H} , the left and right singular vectors are mirror images of each other. This is because the mirror image permutation matrix (a reverse diagonal identity matrix) times any Toeplitz matrix times the transpose permutation matrix equals the original Toeplitz matrix. Let $\mathbf{u}_i(n)$ and $\mathbf{v}_i(n)$ represent the n^{th} vector elements. Then, as verified in (Campbell 1991):

$$\mathbf{u}_i(n) = \pm \mathbf{v}_i(N - n + 1) \quad (\text{III-20})$$

Now, let's consider the special case of an input, \mathbf{x} , equal to one of the right singular vectors, \mathbf{v}_k :

$$\begin{aligned} \mathbf{y} &= \mathbf{H}\mathbf{x} & \mathbf{x} &= \mathbf{v}_k \\ &= \sum_{i=1}^N \sigma_i \mathbf{u}_i \mathbf{v}_i^T \mathbf{v}_k & \mathbf{v}_i^T \mathbf{v}_k &= \delta_{i-k} \\ &= \sigma_k \mathbf{u}_k \end{aligned} \quad (\text{III-21})$$

The output is solely the singular value times the left singular vector corresponding to the input. What if the right singular input is changed by a scale factor α ?

$$\begin{aligned} \mathbf{y} &= \mathbf{H}\mathbf{x} & \mathbf{x} &= \alpha \mathbf{v}_k \\ &= \sum_{i=1}^N \sigma_i \mathbf{u}_i \mathbf{v}_i^T \mathbf{v}_k \alpha & \mathbf{v}_i^T \mathbf{v}_k &= \delta_{i-k} \\ &= \alpha \sigma_k \mathbf{u}_k \end{aligned} \quad (\text{III-22})$$

Thus, a scaling change in one right singular vector at the input produces a change in only one left singular vector at the output. This suggests a method of transform coding.

So, how are the right singular vectors related to our time varying LP filter? Let's do an SVD of $\mathbf{H}^T\mathbf{H}$:

$$\begin{aligned}\mathbf{H}^T\mathbf{H} &= (\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T)^T\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T \\ &= \mathbf{V}\mathbf{\Sigma}^T\mathbf{U}^T\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T \\ &= \mathbf{V}\mathbf{\Sigma}^T\mathbf{\Sigma}\mathbf{V}^T \\ &= \mathbf{V}\mathbf{\Sigma}^2\mathbf{V}^T\end{aligned}\tag{III-23}$$

Thus, this EVD form shows that the right singular vectors, \mathbf{v}_i^T , are also the eigenvectors of $\mathbf{H}^T\mathbf{H}$. The ij element of $\mathbf{H}^T\mathbf{H}$ is:

$$[\mathbf{H}^T\mathbf{H}]_{ij} = \sum_{k=1}^N h_{k-i}h_{k-j}\tag{III-24}$$

$\mathbf{H}^T\mathbf{H}$ is called the *autocorrelation matrix of the impulse response* of the LP filter because it approximates this as N gets large (Atal 1989).

Thus, when the LP filter is varied, we can determine the right singular vectors that will exhibit the desirable localization properties shown above, so now we're ready to consider an SVD basis representation.

SVD Transform Representation

In our case, we'd like to transform the time domain signal to a domain where the effects of changes in the input can be easily controlled in the output. As shown in the previous section, SVD of the filter appears to be an appealing solution.

Let's rewrite the ZSR equation for \mathbf{y} :

$$\mathbf{y} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T\mathbf{x}\tag{III-25}$$

Transform both sides by the left SVD matrix:

$$\mathbf{U}^T \mathbf{y} = \mathbf{\Sigma} \mathbf{V}^T \mathbf{x} \quad (\text{III-26})$$

Now, let's transform the time domain signals to the SVD domain according to:

$$\boldsymbol{\theta} = \mathbf{U}^T \mathbf{y} \quad \boldsymbol{\xi} = \mathbf{V}^T \mathbf{x} \quad (\text{III-27})$$

Thus, the input is represented as a linear combination on the orthonormal basis signals derived from SVD of the matrix of the LP filter's truncated impulse response.

The SVD domain input-output LP filter relation becomes:

$$\boldsymbol{\theta} = \mathbf{\Sigma} \boldsymbol{\xi} \quad (\text{III-28})$$

Or, in scalar notation:

$$\theta_n = \sigma_n \xi_n \quad n = 1, 2, \dots, N \quad (\text{III-29})$$

Thus, by using SVD, we've transformed the LP filter convolution to a simple multiplication that provides significant advantages in interpreting the role of each SV component in the excitation. The singular value components associated with the input and output signals are proportional. Thus, an error in one input component affects only the same output component.

Recall that the singular values, σ_n , are in descending order, so the elements near the beginning of the transformed input vector will dominate the output in the SVD domain. This allows us to determine numerically significant transformed excitation components based upon the size of the singular values. Now, we need to find the *perceptually significant* transformed excitation components.

Perceptually Based SVD

The eigenvectors of the autocorrelation matrix for a typical LP filter for voiced speech demonstrate an approximately sinusoidal shape (Atal 1989). This suggests that the eigenvectors have a narrow-band spectrum, resembling bandpass filters. As shown in Figure III-2, this is verified using the LP coefficients extracted from the vowel /U/. The sinusoidal structure of the right SVs is especially apparent in the early components.

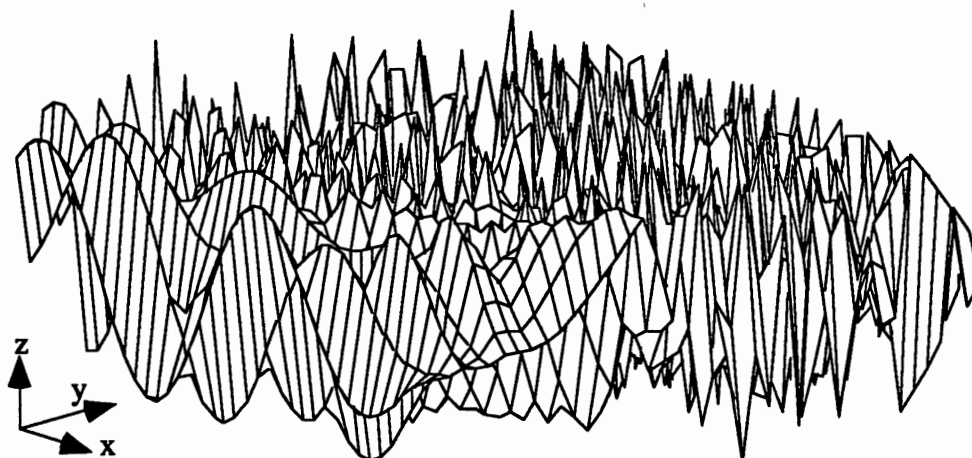


Figure III-2. Sinusoidal Structure of Right SVs of H

As is clear from Figure III-3, the “narrow mountain range” verifies the claim that the right SVs of H have a bandpass characteristic (for the LP filter chosen in this example). Note that the middle right SVs have a bimodal (dual bandpass) characteristic, which also agrees with Atal’s findings.

So, let’s try to perceptually exploit this narrow-band property of the SVD transform domain basis vectors. A very important perceptual process is called *auditory masking*

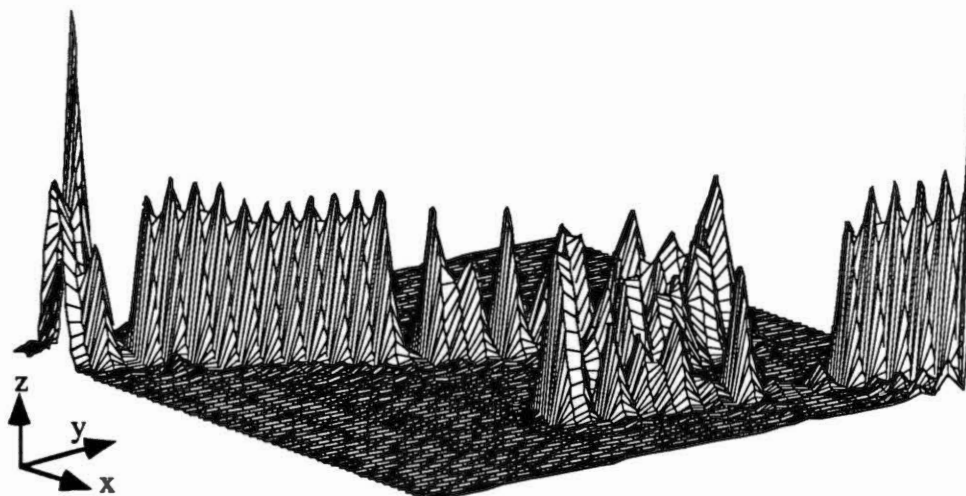


Figure III-3. Narrow-Band Structure of Power Spectrum of Right SVs of H

(Tobias 1970). Auditory masking is the perceptual obscuring of one sound by another. This obscuring occurs when two sounds are in close temporal proximity (i.e., forward/backward masking) or close in frequency proximity (i.e., simultaneous masking).

Critical bands are used to quantify simultaneous masking. Consider a pure tone masked by a band of noise surrounding the tone. As the bandwidth of noise is increased, the amplitude of the noise needed to just mask the tone decreases, but not forever. Beyond a critical bandwidth (which depends on the frequency of the tone), the noise amplitude needed to mask is constant, no matter how wide the noise bandwidth. This frequency versus critical bandwidth relation is known as the *Bark scale* and was depicted in Figure II-7.

The effects of masking of noise by tones not in the same critical band is known as *out-band masking*. It can be quantified by the signal-to-noise ratio (SNR) necessary to mask a critical bandwidth noise placed in other critical bands. For each tone centered in a critical band, the masking threshold (the minimum SNR needed to keep the noise signal masked by the tone) as a function of frequency is reported in reference (De Iacovo and

others 1990). The masking threshold can be determined by (De Iacovo and others 1990) as shown in Table III-3.

TABLE III-3
CALCULATING THE MASKING THRESHOLD

-
- 1) Determine a critical-band piecewise LP spectrum

 - 2) Compute the masking threshold for each band

 - 3) Evaluate the overall noise level at the masking threshold by adding the contributions of each band

Because of the bandpass filter characteristics of the singular vectors and the error component isolation properties of the SVD, it is now possible to determine (for each LP filter) which transformed excitation components control the frequency regions where the ear can tolerate larger errors. This allows us to determine the perceptually significant singular components of the transformed excitation.

Speech Coding

In speech coding, one goal is to reduce the data rate needed to transmit the excitation signal, while providing perceptually high quality synthesized speech. We can do precisely this using the SVD transform domain techniques just presented. One could quantize the excitation components based upon their perceptual significance. The quantization should be adaptive because the perceptual significance of these excitation components is based upon the varying LP filter.

The SVD synthesis equation, $\theta_n = \sigma_n \xi_n$, shows that changes associated with large singular values produce large changes in the output speech. Thus, to minimize distortion, transformed excitation components with large singular values require greater precision in coding relative to components with small singular values. An oversimplified procedure would be to select the largest (first) K transformed excitation components (Sanchez-Calle and others 1990). A better scheme could allocate a different number of bits to different SV components depending upon their singular values. Atal reports that using such a scheme to allow coding of 2 bits per excitation sample on the average is sufficient to keep the quantizing noise inaudible, and, if differential vector quantization is used, this can be reduced to 1 bit (Atal 1989). This is still a rather high data rate, but we have yet to include perceptual knowledge. An even better scheme uses the perceptual threshold masking information discussed above to select the transformed excitation components for transmission. By using this scheme, the number of excitation components can be reduced by 25% without introducing any audible distortion (De Iacovo and others 1990).

Vector Quantization Code Book Design

Vector quantization code books could be designed in the SV transform space. An iterative design procedure was introduced by De Iacovo (De Iacovo and others 1990). Unfortunately, this procedure was overly simplified (e.g., the pitch predictor was omitted) to make it mathematically tractable and computationally feasible. It appears that this area has yet to be fully harvested.

Fast Vector Quantization

SVD could allow a fast vector quantization of the excitation signal (Trancoso and Atal 1990). If the excitation code book is transformed to the SVD domain, a fast search procedure could be based upon only the perceptually significant SV components. The computational cost of transforming the code book for each new linear predictive coding

(LPC) frame might be more compute intensive than traditional, brute force search methods. To avoid the computational expense of transforming a code book for each LP filter, at the expense of memory, one could store multiple code books, with each one corresponding to a class of LP filters.

Speaker Authentication

SVD might provide a compact and powerful set of observation vectors that efficiently capture speaker-dependent features in a perceptually meaningful sense. Vector quantization (VQ) based speaker authentication approaches, as shown later in Equation V-4, could benefit from the VQ code book design and fast VQ ideas outlined above. The perceptual aspects of SVD could be exploited to yield perceptually meaningful features in speaker authentication applications. Use of SVD in speaker authentication has not been reported in the literature. The speaker discriminatory power of SVD based features was investigated.

As shown previously in Figure III-2, the response of the singular vectors have desirable properties closely related to perception. Since the response of the singular vectors is not directly controllable, using them as features was not further pursued at this time. However, in the process of feature selection, the power of SVD was brought to bear on the analysis of poorly conditioned covariance matrices.

SVD Advantages

The SVD allows representation of perceptually based errors in a transform domain where each error on the transformed excitation signal is reflected only in the same component of the transformed output signal. Perceptual phenomena is accounted for by the band-limited characteristics of the singular vectors used in the decomposition. This allows us to reduce the number of transformed excitation components in a systematic and perceptually meaningful way. This reduction in the transformed excitation components is

useful for efficient coding, fast vector quantization, and vector quantization code book design.

Generalized SVD

The basic SVD utilizes *one* matrix and the singular values can be used for various applications as outlined above. We can still use this approach and more by using the generalized singular value decomposition (GSVD) which uses *two* matrices.

Given two matrices, A and B:

$$\mathbf{A} \in \mathfrak{R}^{m \times n} \quad (m \geq n) \quad \mathbf{B} \in \mathfrak{R}^{p \times n} \quad (\text{III-30})$$

There exist orthogonal matrices, $\mathbf{U} \in \mathfrak{R}^{m \times m}$ and $\mathbf{V} \in \mathfrak{R}^{p \times p}$, and an invertible $\mathbf{X} \in \mathfrak{R}^{n \times n}$ such that:

$$\begin{aligned} \mathbf{U}^T \mathbf{A} \mathbf{X} &= \mathbf{D}_A = \text{diag}(\alpha_1, \alpha_2, \dots, \alpha_n), \quad \alpha_i \geq 0 \\ \mathbf{V}^T \mathbf{B} \mathbf{X} &= \mathbf{D}_B = \text{diag}(\beta_1, \beta_2, \dots, \beta_q), \quad \beta_i \geq 0; \quad q = \min\{p, n\} \end{aligned} \quad (\text{III-31})$$

where:

$$\beta_1 \geq \dots \geq \beta_r > \beta_{r+1} = \dots = \beta_q = 0 \quad r = \text{rank}(\mathbf{B}) \quad (\text{III-32})$$

This is the *generalized singular value decomposition* of A and B (Golub and Van Loan 1983). The simultaneous equations are coupled through X. Since X need not be a unitary matrix (just a nonsingular matrix), it may have some parameters that could give some normalization and other information.

If A and B are sequential measurements of a speech signal, then X might contain information about transition regions between A and B. GSVD could be used on sequences of speech measurement matrices to yield an \mathbf{X}_i sequence of matrices. These \mathbf{X}_i

matrices could be used as features in a speaker authentication system as a means to capture transition information (Yarlagadda 1991).

Reproducing accurate transition regions of speech are crucial to its high quality synthesis. The speaker discriminatory power of speech transition regions using GSVD was briefly evaluated. Normalization problems precluded this from being useful at the present time. Hopefully, these problems will be solved in the future.

The next chapter presents feature selection, estimation of mean and covariance, divergence, and Bhattacharyya distance. It is highlighted by the development of the divergence shape measure and the Bhattacharyya distance shape.

CHAPTER IV

FEATURE SELECTION AND MEASURES

In order to apply mathematical tools, and without loss of generality, the speech signal can be represented by a sequence of feature vectors. In this section, the selection of appropriate features is discussed along with methods to estimate (extract or measure) them. This is known as feature selection and feature extraction.

Traditionally, pattern recognition paradigms are divided into three components: feature extraction and selection, pattern matching, and classification. Although this division is convenient from the perspective of designing system components, these components are not independent. The false demarcation among these components can lead to suboptimal designs because they all interact in real-world systems.

In speaker authentication, the goal is to design a system that minimizes the probability of authentication errors. Thus, the underlying objective is to discriminate between the given speaker and all others. A modern comprehensive review of the state of the art in discriminant analysis is given in (Gnanadesikan and Kettenring 1989).

Traditional Feature Selection

Feature extraction is the estimation (measurement) of variables, called an observation vector, from another set of variables (e.g., a speech signal time series). Feature selection is the transformation of these observation vectors to feature vectors. The goals of feature selection are to find a transformation that preserves the information pertinent to the application, to realize a transform domain where meaningful comparisons can be

performed using simple distance measures, and to form a relatively low dimensional feature space.

Although it might be tempting at first to select all the extracted features, the “curse of dimensionality” quickly becomes overwhelming (Duda and Hart 1973). As more features are used, the feature dimensions increase, which imposes severe requirements on computation and storage in both training and testing. The demand for a large number of training samples grows exponentially with the dimension of the feature space. This severely restricts the usefulness of nonparametric procedures (no assumed underlying statistical model) and nonlinear transforms because this compounds their voracious appetite for large training sets.

The traditional statistical methods to reduce dimensionality, and avoid this curse, are principal component analysis and factor analysis. Principal component analysis seeks to find a lower dimensional representation that accounts for variance of the features. Factor analysis seeks to find a lower dimensional representation that accounts for correlations among the features. In other disciplines, principal component analysis is called the *Karhunen-Loève expansion* (KLE) or *eigenvector orthonormal expansion*. Since each eigenvector can be ranked by its corresponding eigenvalue, a subset of the eigenvectors can be chosen to minimize the mean square error in representing the data. Although KLE is optimum for representing classes with the same mean, it is not necessarily optimum for discriminating between classes (Tou and Gonzalez 1974). Since speaker authentication is a discrimination problem instead of a representation problem, we seek other means to reduce the dimensionality of the data.

Linear transformations are capable of dividing the feature space by a hyperplane. If data is *linearly separable*, then it can be discriminated by a hyperplane. In the case of a two-dimensional feature space, the hyperplane collapses to a line. As shown below, if $p(\mathbf{x}) \sim N(\boldsymbol{\mu}_x, \mathbf{C}_x)$, \mathbf{A} is an m by n matrix, and $\mathbf{y} = \mathbf{A}\mathbf{x}$ is an m -component image vector, then $p(\mathbf{y}) \sim N(\mathbf{A}\boldsymbol{\mu}_x, \mathbf{A}\mathbf{C}_x\mathbf{A}^T)$.

$$\begin{aligned}
\mathbf{y} &= \mathbf{Ax} \\
\boldsymbol{\mu}_y &= E[\mathbf{y}] = E[\mathbf{Ax}] = \mathbf{A}E[\mathbf{x}] \\
&= \mathbf{A}\boldsymbol{\mu}_x \\
\mathbf{C}_y &= E[(\mathbf{y} - \boldsymbol{\mu}_y)(\mathbf{y} - \boldsymbol{\mu}_y)^T] = E[\mathbf{A}(\mathbf{x} - \boldsymbol{\mu}_x)(\mathbf{A}(\mathbf{x} - \boldsymbol{\mu}_x))^T] \\
&= E[\mathbf{A}(\mathbf{x} - \boldsymbol{\mu}_x)(\mathbf{x} - \boldsymbol{\mu}_x)^T \mathbf{A}^T] = \mathbf{A}E[(\mathbf{x} - \boldsymbol{\mu}_x)(\mathbf{x} - \boldsymbol{\mu}_x)^T] \mathbf{A}^T \\
&= \mathbf{A}\mathbf{C}_x\mathbf{A}^T
\end{aligned} \tag{IV-1}$$

Thus, a linear transformation of a multivariate normal vector also has a normal density. Any linear combination of normally distributed random variables is again normal. This can be used to tremendous advantage if the feature densities of the speakers are assumed normal. This allows us to lump all the other speaker probability density functions (pdf's) into a single normal pdf. Pairwise (two class) discriminators are usually much easier to design than multiclass discriminators. Thus, pairwise discriminators can be designed for the claimant talker versus all other talkers.

In the special case where the transformation is a unit length vector, \mathbf{a} , $y = \mathbf{ax}$ is a scalar that represents the projection of \mathbf{x} onto a line in the direction of \mathbf{a} . In general, $\mathbf{A}\mathbf{C}_x\mathbf{A}^T$ is the variance of the projection of \mathbf{x} onto the column space of \mathbf{A} . Thus, knowledge of the covariance matrix allows us to calculate the dispersion of the data in any direction.

In Figure IV-1, two classes are represented by boxes and circles in a two-dimensional feature space (x_1, x_2) . Here we see that if feature x_1 or x_2 was used by itself, discrimination errors would occur because of the overlap between the projected classes onto the x_1 or x_2 axes. However, it is quite clear that the data is perfectly linearly separable by the dashed line. If the data is linearly transformed onto the column space of \mathbf{A} , perfect discrimination is achieved. In addition, one can see a clustering effect by the reduced variance of the projection onto the column space of \mathbf{A} .

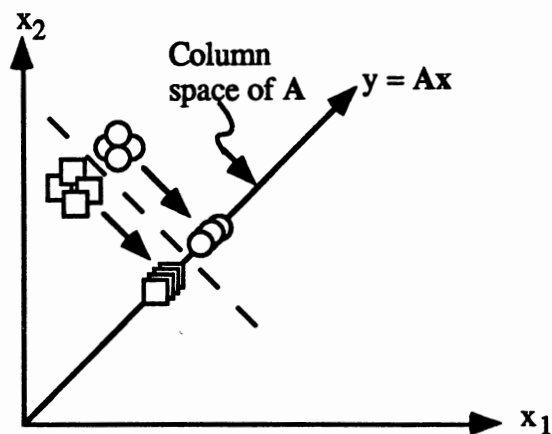


Figure IV-1. Linear Transformation

It should be noted that data may not always be discriminated well by linear transformation. In these cases, nonlinear transformation may lead to improved discrimination. An example of this are the classes defined by the members of interlocking spirals. No line can separate the spirals, but a nonlinear transformation could yield perfect discrimination.

The goal of speaker authentication feature selection is to find a set that minimizes probability of error. Unfortunately, an explicit mathematical expression is unavailable, except for trivial cases, which hinders rigorous mathematical development. Even for normal pdf's, a numerical integration is required to determine probability of error (except for the equal covariance case) (Fukunaga 1990).

To make the problem mathematically tractable, consider discriminant feature selection with respect to a Bayes classifier. This reduces discrimination to the probability of error due to a Bayes classifier. This implies a set that exhibits low intraspeaker variability and high interspeaker variability. A technique that can be used to find good features is analysis of variance (ANOVA), which involves measuring Fisher's F-ratio

(Equation IV-2) between the sample pdf's of different features. For speaker verification, high F-ratios are desirable.

$$F = \frac{\text{variance of speaker means}}{\text{average intraspeaker variance}} \quad (\text{IV-2})$$

Unfortunately, ANOVA requires evaluating the F-ratio for many different *combinations* of features to really be useful. For example, two features with high individual F-ratios might contain redundant information and as a feature vector be less effective than two features which individually had low F-ratios. The usefulness of the F-ratio as a discrimination measure is further reduced if the classes are multimodal or if they have the same means. This is a fatal flaw with any criterion that is dominated by differences between class means. This will now be demonstrated.

Normal Density With Equal Means

The normal pdf is often a good approximation to real-world density functions. Classes will exhibit normal densities when each pattern of a class is a random vector formed by superposition of a random vector upon a nonrandom vector, where the superimposed random vectors are drawn from the same normal density. This is a good approximation to real-world situations characterized by independent identically distributed (i.i.d.) additive Gaussian noise (AGN). The normal pdf has some striking advantages. It is one of the simplest parametric models, being characterized by a mean and variance. In addition, the sum of normal random variables yields a normal random variable.

The *n-variate normal pdf* is defined as:

$$p(\mathbf{x}) = (2\pi)^{-n/2} |\mathbf{C}|^{-1/2} \exp\left[-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \mathbf{C}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right] \quad (\text{IV-3})$$

$$\sim N(\boldsymbol{\mu}, \mathbf{C})$$

where \mathbf{C} is the n -by- n covariance matrix and $\boldsymbol{\mu}$ is an n -dimensional column component mean vector. Note that in Equation IV-3, contours of constant probability occur for values of \mathbf{x} where the argument of the exponential is constant. Neglecting the scale factor, the argument of the exponential is referred to as the *Mahalanobis distance*, d_M^2 , between \mathbf{x} and $\boldsymbol{\mu}$:

$$d_M^2 = (\mathbf{x} - \boldsymbol{\mu})^T \mathbf{C}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \quad (\text{IV-4})$$

Thus, the loci of points of constant density are hyperellipsoids of constant Mahalanobis distance to $\boldsymbol{\mu}$. The principal axes of these hyperellipsoids are given by the eigenvectors of \mathbf{C} and their eigenvalues determine the lengths of the corresponding axes.

Samples drawn from a multivariate normal density tend to cluster. The center of the cluster is determined by the mean and the shape of the cluster is determined by the covariance matrix. In the bivariate ($n=2$) case, it is convenient for purposes of display to show the 1-sigma ellipse. For example, Figure IV-2 shows the bivariate 1-sigma ellipses for two classes with equal means, $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = [0 \ 0]$, and unequal covariance matrices.

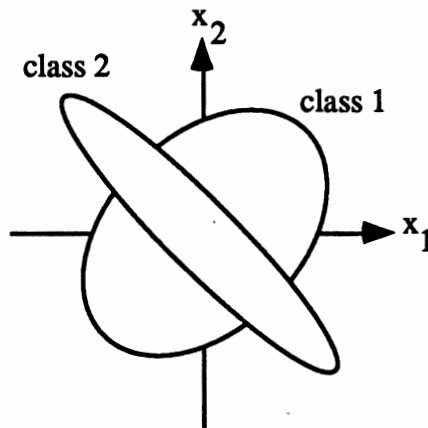


Figure IV-2. Unequal Covariance

Although there is no line that can perfectly discriminate these two classes, it's easy to visualize that a 45 degree projection would provide some discrimination power. However, the F-ratio would indicate that these features, x_1 and x_2 , are powerless because the classes have the same means in the x_1-x_2 space.

Now consider a bimodal pdf. Figure IV-3 shows class 1 as being bimodal in x_1 . The means of both classes are the same; hence, the F-ratio would show feature x_1 as powerless. However, it is clear from Figure IV-3 that x_1 is powerful because significant discriminatory information exists along feature x_1 .

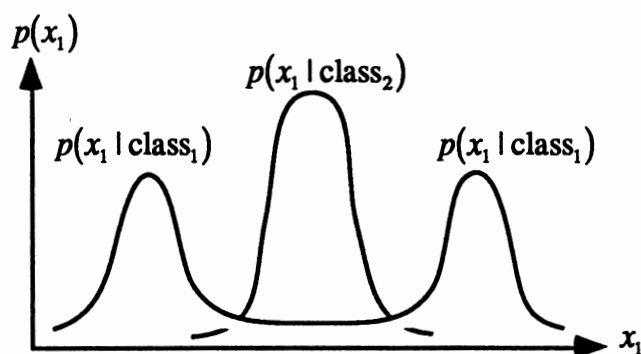


Figure IV-3. A Bimodal Class

Thus, caution should be used with any criteria, such as the F-ratio, that relies on class means. If the classes have the same means or are not unimodal, the F-ratio can be a poor measure of discrimination power. Clearly, we seek a criterion that more accurately portrays discrimination power.

Importance Sampling

Statistical methods such as *importance sampling* (Whalen 1971) are used to study outliers (e.g., wolves and sheep). Importance sampling relies on models of the outlier pdf's, which are unknown for wolves and sheep. Importance sampling may allow systematic study of wolves and sheep. Perhaps the parameters necessary for accurate importance sampling models could be derived from the massive YOHO database.

Mean and Covariance Estimation

The unbiased estimate (UBE) of the covariance is given by the sample covariance:

$$\hat{\mathbf{C}} = \frac{1}{N-1} \sum_{i=1}^N (\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^T \quad (\text{IV-5})$$

The UBE and maximum likelihood estimates (MLE) of covariance differ by their scaling factors of $\frac{1}{N-1}$ and $\frac{1}{N}$, respectively. They are both termed a sample covariance matrix. When the mean is being estimated too, the UBE is generally preferred; however, they are practically identical when N is large.

To estimate the mean and covariance when all samples are not yet available or when dealing with a large number of samples, recursive computation methods are desirable. Denoting an estimate based upon N samples as $\hat{\boldsymbol{\mu}}_N$ and on N+1 samples as $\hat{\boldsymbol{\mu}}_{N+1}$, the sample mean is:

$$\begin{aligned} \hat{\boldsymbol{\mu}}_{N+1} &= \frac{1}{N+1} \sum_{k=1}^{N+1} \mathbf{x}_k \\ &= \frac{1}{N+1} \left(\sum_{k=1}^N \mathbf{x}_k + \mathbf{x}_{N+1} \right) \\ &= \frac{1}{N+1} (N\hat{\boldsymbol{\mu}}_N + \mathbf{x}_{N+1}) \\ &= \hat{\boldsymbol{\mu}}_N + \frac{1}{N+1} (\mathbf{x}_{N+1} - \hat{\boldsymbol{\mu}}_N) \end{aligned} \quad (\text{IV-6})$$

The UBE sample covariance matrix recursion can be similarly derived, with $\hat{\mathbf{C}}_N$ representing the estimate based upon N samples:

$$\begin{aligned}
 \hat{\mathbf{C}}_N &= \frac{1}{N-1} \sum_{k=1}^N (\mathbf{x}_k - \hat{\boldsymbol{\mu}}_N)(\mathbf{x}_k - \hat{\boldsymbol{\mu}}_N)^T \\
 &= \frac{1}{N-1} \left[\left(\sum_{k=1}^N \mathbf{x}_k \mathbf{x}_k^T \right) - N \hat{\boldsymbol{\mu}}_N \hat{\boldsymbol{\mu}}_N^T \right] \\
 \hat{\mathbf{C}}_{N+1} &= \frac{1}{N} \sum_{k=1}^{N+1} (\mathbf{x}_k - \hat{\boldsymbol{\mu}}_{N+1})(\mathbf{x}_k - \hat{\boldsymbol{\mu}}_{N+1})^T \\
 &= \frac{1}{N} \left[\left(\sum_{k=1}^{N+1} \mathbf{x}_k \mathbf{x}_k^T \right) - (N+1) \hat{\boldsymbol{\mu}}_{N+1} \hat{\boldsymbol{\mu}}_{N+1}^T \right] \\
 &= \frac{1}{N} \left[\left(\sum_{k=1}^N \mathbf{x}_k \mathbf{x}_k^T + \mathbf{x}_{N+1} \mathbf{x}_{N+1}^T \right) - (N+1) \hat{\boldsymbol{\mu}}_{N+1} \hat{\boldsymbol{\mu}}_{N+1}^T \right] \\
 &= \frac{N-1}{N} \hat{\mathbf{C}}_N + \hat{\boldsymbol{\mu}}_N \hat{\boldsymbol{\mu}}_N^T + \frac{1}{N} \mathbf{x}_{N+1} \mathbf{x}_{N+1}^T - \frac{N+1}{N} \hat{\boldsymbol{\mu}}_{N+1} \hat{\boldsymbol{\mu}}_{N+1}^T \\
 &= \frac{N-1}{N} \hat{\mathbf{C}}_N + \frac{1}{N+1} (\mathbf{x}_{N+1} - \hat{\boldsymbol{\mu}}_N)(\mathbf{x}_{N+1} - \hat{\boldsymbol{\mu}}_N)^T
 \end{aligned} \tag{IV-7}$$

Sample covariance matrices using LSP features are shown in the mesh plots of Figures IV-4 and IV-5. In each plot, the variances and covariances of 10 LSP coefficients are represented in the vertical direction on a 10 x 10 mesh. From a total of 80 seconds of speech, each matrix (mesh plot) was generated from the LSP vectors corresponding to voiced speech.

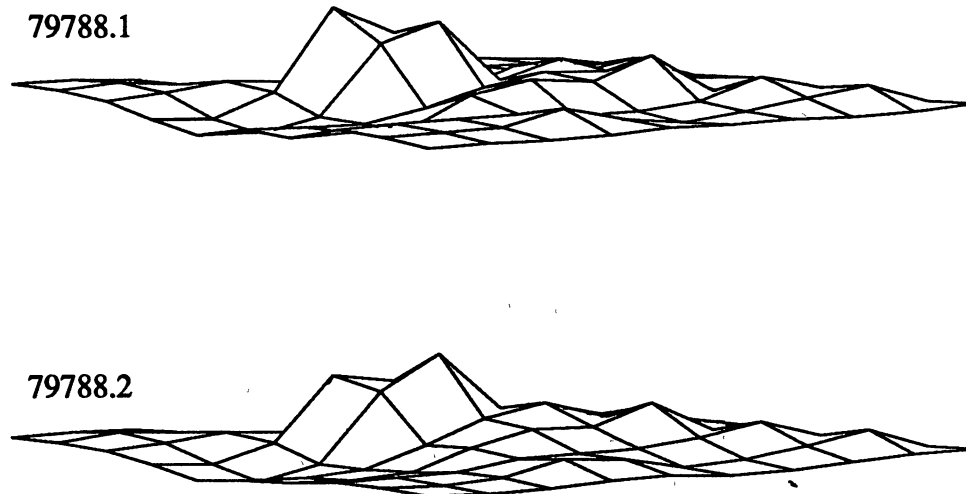


Figure IV-4. LSP Covariance Matrices: Different Sessions, Same Speaker

Notice that these covariance matrices for different sessions of the same speaker appear to be similar.

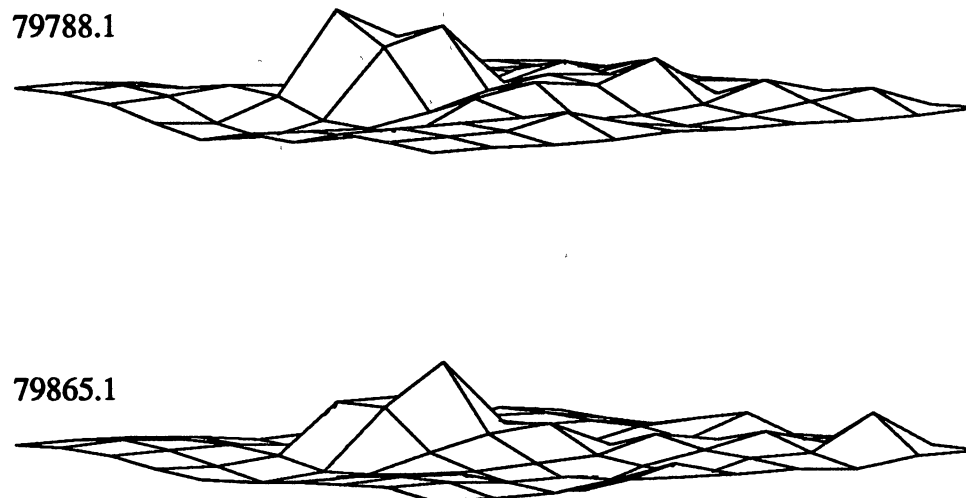


Figure IV-5. LSP Covariance Matrices, Different Speakers

These LSP covariance matrices appear to have more differences between speakers than similarities for the same speaker. As shown later, the LSP covariance matrices can capture speaker identity.

Divergence Measure

Divergence is a measure of distance or dissimilarity between two classes based upon information theory (Kullback 1968). It provides a means of feature ranking and evaluation of class discrimination effectiveness (Tou and Gonzalez 1974). The following equations are based upon Tou's rather complicated derivation (Tou and Gonzalez 1974). To allow the reader to more readily understand Tou's equations, the derivations are given in easy to follow form; complete with intermediate steps. Let the probability of occurrence of pattern \mathbf{x} , given that it belongs to class ω_i , be:

$$p_i(\mathbf{x}) = p(\mathbf{x} | \omega_i) \quad (\text{IV-8})$$

and likewise for class ω_j :

$$p_j(\mathbf{x}) = p(\mathbf{x} | \omega_j) \quad (\text{IV-9})$$

Then, the *discriminating information* of an observation \mathbf{x} , in the Bayes classifier sense, for class ω_i versus class ω_j , can be measured by the logarithm of the likelihood ratio:

$$u_{ij} = \ln \frac{p_i(\mathbf{x})}{p_j(\mathbf{x})} \quad (\text{IV-10})$$

Entropy is the statistical measure of information or uncertainty. The *population entropy* for a given ensemble of pattern vectors is:

$$H = -E[\ln p(\mathbf{x})] \quad (\text{IV-11})$$

The entropy of the i^{th} population of patterns is:

$$H_i = - \int_{\mathbf{x}} p_i(\mathbf{x}) \ln p_i(\mathbf{x}) d\mathbf{x} \quad (\text{IV-12})$$

The *average discriminating information* for class ω_i versus class ω_j , over all observations, also known as *directed divergence*, *Kullback-Leibler number* (Kullback 1968), or *discrimination* (Blahut 1987), is then:

$$\begin{aligned} I(i, j) &= \int_{\mathbf{x}} p_i(\mathbf{x}) u_{ij} d\mathbf{x} \\ &= \int_{\mathbf{x}} p_i(\mathbf{x}) \ln \frac{p_i(\mathbf{x})}{p_j(\mathbf{x})} d\mathbf{x} \end{aligned} \quad (\text{IV-13})$$

Likewise, the discriminating information for class ω_j versus class ω_i , can be measured by the logarithm of the likelihood ratio:

$$u_{ji} = \ln \frac{p_j(\mathbf{x})}{p_i(\mathbf{x})} \quad (\text{IV-14})$$

The average discriminating information for class ω_j is then:

$$I(j, i) = \int_{\mathbf{x}} p_j(\mathbf{x}) \ln \frac{p_j(\mathbf{x})}{p_i(\mathbf{x})} d\mathbf{x} \quad (\text{IV-15})$$

The *divergence* (the symmetric directed divergence) is defined as the total average information for discriminating class ω_i from class ω_j :

$$\begin{aligned} J_{ij} &= I(i, j) + I(j, i) \\ &= \int_{\mathbf{x}} [p_i(\mathbf{x}) - p_j(\mathbf{x})] \ln \frac{p_i(\mathbf{x})}{p_j(\mathbf{x})} d\mathbf{x} \end{aligned} \quad (\text{IV-16})$$

Now, in order to select features with this measure, we need the feature pdf for each pattern class. Assuming the pattern classes are n-variate normal populations:

$$p_i(\mathbf{x}) \sim N(\boldsymbol{\mu}_i, \mathbf{C}_i) \quad p_j(\mathbf{x}) \sim N(\boldsymbol{\mu}_j, \mathbf{C}_j) \quad (\text{IV-17})$$

Substituting Equation IV-3 into Equation IV-10 yields the log likelihood ratio:

$$u_y = \frac{1}{2} \ln \frac{|\mathbf{C}_j|}{|\mathbf{C}_i|} - \frac{1}{2} \text{tr}[\mathbf{C}_i^{-1}(\mathbf{x} - \boldsymbol{\mu}_i)(\mathbf{x} - \boldsymbol{\mu}_i)^T] + \frac{1}{2} \text{tr}[\mathbf{C}_j^{-1}(\mathbf{x} - \boldsymbol{\mu}_j)(\mathbf{x} - \boldsymbol{\mu}_j)^T] \quad (\text{IV-18})$$

The average information for discrimination between these two classes is:

$$\begin{aligned} I(i, j) &= \int_{\mathbf{x}} p_i(\mathbf{x}) u_y d\mathbf{x} \\ &= \int_{\mathbf{x}} (2\pi)^{-n/2} |\mathbf{C}_i|^{-1/2} \exp\left[-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_i)^T \mathbf{C}_i^{-1}(\mathbf{x} - \boldsymbol{\mu}_i)\right] \cdot \\ &\quad \left\{ \frac{1}{2} \ln \frac{|\mathbf{C}_j|}{|\mathbf{C}_i|} - \frac{1}{2} \text{tr}[\mathbf{C}_i^{-1}(\mathbf{x} - \boldsymbol{\mu}_i)(\mathbf{x} - \boldsymbol{\mu}_i)^T] + \right. \\ &\quad \left. \frac{1}{2} \text{tr}[\mathbf{C}_j^{-1}(\mathbf{x} - \boldsymbol{\mu}_j)(\mathbf{x} - \boldsymbol{\mu}_j)^T] \right\} d\mathbf{x} \\ &= \frac{1}{2} \ln \frac{|\mathbf{C}_j|}{|\mathbf{C}_i|} + \frac{1}{2} \text{tr}[\mathbf{C}_i(\mathbf{C}_j^{-1} - \mathbf{C}_i^{-1})] + \frac{1}{2} \text{tr}[\mathbf{C}_j^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^T] \end{aligned} \quad (\text{IV-19})$$

Let the difference in the means be represented as:

$$\boldsymbol{\delta} = \boldsymbol{\mu}_i - \boldsymbol{\mu}_j \quad (\text{IV-20})$$

The average information for discrimination between these two classes is:

$$I(i, j) = \frac{1}{2} \ln \frac{|\mathbf{C}_j|}{|\mathbf{C}_i|} + \frac{1}{2} \text{tr}[\mathbf{C}_i(\mathbf{C}_j^{-1} - \mathbf{C}_i^{-1})] + \frac{1}{2} \text{tr}[\mathbf{C}_j^{-1} \boldsymbol{\delta} \boldsymbol{\delta}^T] \quad (\text{IV-21})$$

Hence, the divergence for these two normally distributed classes is:

$$\begin{aligned}
J_{ij} &= \frac{1}{2} \ln \frac{|C_j|}{|C_i|} + \frac{1}{2} \text{tr}[C_i(C_j^{-1} - C_i^{-1})] + \frac{1}{2} \text{tr}[C_j^{-1}(\mu_i - \mu_j)(\mu_i - \mu_j)^T] \\
&\quad + \frac{1}{2} \ln \frac{|C_i|}{|C_j|} + \frac{1}{2} \text{tr}[C_j(C_i^{-1} - C_j^{-1})] + \frac{1}{2} \text{tr}[C_i^{-1}(\mu_j - \mu_i)(\mu_j - \mu_i)^T] \quad (\text{IV-22}) \\
&= \frac{1}{2} \text{tr}[(C_i - C_j)(C_j^{-1} - C_i^{-1})] + \frac{1}{2} \text{tr}[(C_i^{-1} + C_j^{-1})(\mu_i - \mu_j)(\mu_i - \mu_j)^T] \\
&= \frac{1}{2} \text{tr}[(C_i - C_j)(C_j^{-1} - C_i^{-1})] + \frac{1}{2} \text{tr}[(C_i^{-1} + C_j^{-1})\delta\delta^T]
\end{aligned}$$

Note that Equation IV-22 is the sum of two components, one is based solely upon differences between the covariance matrices and the other involves differences between the mean vectors, δ . These components can be characterized respectively as differences in shape and size of the pdf's. This shape component, the *divergence shape*, will prove very useful later on:

$$J'_{ij} = \text{tr}[(C_i - C_j)(C_j^{-1} - C_i^{-1})] \quad (\text{IV-23})$$

Equation IV-22 is slightly complicated, so let us consider two simplifying special cases.

Equal Covariance Divergence

First, for the equal covariance case, let:

$$C_i = C_j = C \quad (\text{IV-24})$$

This leaves only the last term from Equation IV-19:

$$\begin{aligned}
I(i, j) &= \frac{1}{2} \text{tr}[C^{-1}(\mu_i - \mu_j)(\mu_i - \mu_j)^T] \\
&= \frac{1}{2} \text{tr}[C^{-1}\delta\delta^T] \quad (\text{IV-25}) \\
&= \frac{1}{2} \delta^T C^{-1} \delta
\end{aligned}$$

and, therefore:

$$\begin{aligned}
 J_{ij} &= \frac{1}{2} \text{tr}[\mathbf{C}^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^T] + \frac{1}{2} \text{tr}[\mathbf{C}^{-1}(\boldsymbol{\mu}_j - \boldsymbol{\mu}_i)(\boldsymbol{\mu}_j - \boldsymbol{\mu}_i)^T] \\
 &= \text{tr}[\mathbf{C}^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^T] \\
 &= \boldsymbol{\delta}^T \mathbf{C}^{-1} \boldsymbol{\delta}
 \end{aligned} \tag{IV-26}$$

Comparing this with Equation IV-4, the divergence for this normal equal covariance case is simply the Mahalanobis distance between the two class means.

For a univariate ($n=1$) normal equal variance, σ^2 , population:

$$I(i, j) = \frac{1}{2} \frac{(\mu_i - \mu_j)^2}{\sigma^2} \tag{IV-27}$$

Reassuringly, the divergence in this equal covariance case is the familiar F-ratio:

$$J_{ij} = \frac{(\mu_i - \mu_j)^2}{\sigma^2} \tag{IV-28}$$

Equal Mean Divergence

Next, for the equal population means case:

$$\boldsymbol{\mu}_i = \boldsymbol{\mu}_j, \quad \boldsymbol{\delta} = 0 \tag{IV-29}$$

The average information is:

$$\begin{aligned}
 I(i, j) &= \frac{1}{2} \ln \frac{|\mathbf{C}_j|}{|\mathbf{C}_i|} + \frac{1}{2} \text{tr}[\mathbf{C}_i(\mathbf{C}_j^{-1} - \mathbf{C}_i^{-1})] \\
 &= \frac{1}{2} \ln \frac{|\mathbf{C}_j|}{|\mathbf{C}_i|} + \frac{1}{2} \text{tr}[\mathbf{C}_i \mathbf{C}_j^{-1}] - \frac{n}{2}
 \end{aligned} \tag{IV-30}$$

The divergence is:

$$\begin{aligned}
J_{ij} &= \frac{1}{2} \text{tr}[(C_i - C_j)(C_j^{-1} - C_i^{-1})] \\
&= \frac{1}{2} \text{tr}[C_i C_j^{-1}] + \text{tr}[C_j C_i^{-1}] - n
\end{aligned}
\tag{IV-31}$$

Divergence Properties

The divergence satisfies all the metric properties except the triangle inequality; thus, divergence is not termed a distance (Kullback and Leibler 1951). The following properties of divergence are proven in the landmark paper of Kullback and Leibler (Kullback and Leibler 1951). Positivity (i.e., almost positive definite) and symmetry properties are satisfied:

$$\begin{aligned}
J_{ij} &\geq 0 \quad \text{and} \quad J_{ij} = 0 \text{ iff } p_i \neq p_j \\
J_{ij} &= J_{ji}
\end{aligned}
\tag{IV-32}$$

By counterexample, divergence can be shown to violate the triangle inequality by taking $p_1 \sim N(0,1)$, $p_2 \sim N(0,4)$, and $p_3 \sim N(0,5)$; thus, $J_{13} > J_{12} + J_{23}$.

Additional measurements (increased dimensionality) cannot decrease divergence:

$$J_{ij}(x_1, x_2, \dots, x_m) \leq J_{ij}(x_1, x_2, \dots, x_m, x_{m+1})
\tag{IV-33}$$

As should be expected from an information-theoretic measure, processing cannot increase divergence (Blahut 1987). Thus, transformation of the feature space must maintain or decrease divergence. Furthermore, divergence can be shown to be invariant under *onto* measurable transformation (Kullback and Leibler 1951). Kullback's real analysis based proof is rather difficult to follow, so let's consider the special case of proving the invariance of the divergence measure under nonsingular linear transformation (affine transformation could be similarly shown):

if $p(x) \sim N(\mu_x, C_x)$ where $x \in \mathcal{R}^n$ and $A \in \mathcal{R}^{m \times n}$

let $y = Ax$ where $y \in \mathcal{R}^m$

then $\mu_y = E[y] = E[Ax] = AE[x] = A\mu_x$

$$C_y = E[(y - \mu_y)(y - \mu_y)^T] = E[(Ax - A\mu_x)(Ax - A\mu_x)^T] = AC_xA^T$$

$$\therefore p(y) \sim N(A\mu_x, AC_xA^T)$$

$$\begin{aligned} \text{let } J_y^{(x)} &= \frac{1}{2} \text{tr} \left[(C_i^{(x)} - C_j^{(x)}) \left((C_j^{(x)})^{-1} - (C_i^{(x)})^{-1} \right) \right] \\ &\quad + \frac{1}{2} \text{tr} \left[\left((C_i^{(x)})^{-1} + (C_j^{(x)})^{-1} \right) (\mu_i^{(x)} - \mu_j^{(x)}) (\mu_i^{(x)} - \mu_j^{(x)})^T \right] \end{aligned}$$

$$\begin{aligned} \text{then } J_y^{(y)} &= \frac{1}{2} \text{tr} \left[(AC_i^{(x)}A^T - AC_j^{(x)}A^T) \right. \\ &\quad \left. \cdot \left((A^T)^{-1}(C_j^{(x)})^{-1}A^{-1} - (A^T)^{-1}(C_i^{(x)})^{-1}A^{-1} \right) \right] \\ &\quad + \frac{1}{2} \text{tr} \left[\left((A^T)^{-1}(C_i^{(x)})^{-1}A^{-1} + (A^T)^{-1}(C_j^{(x)})^{-1}A^{-1} \right) \right. \\ &\quad \left. \cdot (A\mu_i^{(x)} - A\mu_j^{(x)}) (A\mu_i^{(x)} - A\mu_j^{(x)})^T \right] \\ &= \frac{1}{2} \text{tr} \left[A(C_i^{(x)} - C_j^{(x)})A^T (A^T)^{-1} \left((C_j^{(x)})^{-1} - (C_i^{(x)})^{-1} \right) A^{-1} \right] \\ &\quad + \frac{1}{2} \text{tr} \left[(A^T)^{-1} \left((C_i^{(x)})^{-1} + (C_j^{(x)})^{-1} \right) A^{-1} A (\mu_i^{(x)} - \mu_j^{(x)}) \right. \\ &\quad \left. \cdot (A(\mu_i^{(x)} - \mu_j^{(x)}))^T \right] \\ &= \frac{1}{2} \text{tr} \left[AA^{-1}(C_i^{(x)} - C_j^{(x)}) \left((C_j^{(x)})^{-1} - (C_i^{(x)})^{-1} \right) \right] \tag{IV-34} \\ &\quad + \frac{1}{2} \text{tr} \left[(A^T)^{-1} A^T \left((C_i^{(x)})^{-1} + (C_j^{(x)})^{-1} \right) (\mu_i^{(x)} - \mu_j^{(x)}) (\mu_i^{(x)} - \mu_j^{(x)})^T \right] \\ &= J_y^{(x)} \end{aligned}$$

This is a powerful result because of the many useful linear transformations (e.g., discrete Fourier transform, discrete cosine transform, and discrete convolution). For example, if the frequency domain can be attained via linear transformation, there is no need to separately consider this mapping of the features. This invariance also implies that linear feature selection is unnecessary unless dimensionality reduction is desired.

Divergence is additive for independent measurements:

$$J_j(x_1, x_2, \dots, x_m) = \sum_{k=1}^m J_j(x_k) \quad (\text{IV-35})$$

This allows ranking the importance of each feature according to its associated divergence, as shown in the following example.

Example of Equal Mean Divergence

The preceding concepts are demonstrated here based upon an example taken from Tou and Gonzalez (Tou and Gonzalez 1974). Intermediate steps have been added to aid the reader. Given the following observations:

$$\begin{aligned} \mathbf{x}_{11} &= \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} & \mathbf{x}_{12} &= \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} & \mathbf{x}_{13} &= \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix} & \mathbf{x}_{14} &= \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} \\ \mathbf{x}_{21} &= \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} & \mathbf{x}_{22} &= \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} & \mathbf{x}_{23} &= \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix} & \mathbf{x}_{24} &= \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \end{aligned} \quad (\text{IV-36})$$

where the first index indicates class ω_1 or ω_2 . These patterns are shown in Figure IV-6. From this figure, it is obvious that the data could be perfectly discriminated by a plane slicing through the data. Let us see how the divergence metric cuts the data.

To estimate the population means, we approximate the mean vectors by the sample average over N samples:

$$\begin{aligned} \boldsymbol{\mu} &= E[\mathbf{x}] \\ &= \int_{\mathbf{x}} \mathbf{x} p(\mathbf{x}) d\mathbf{x} \\ &\approx \frac{1}{N} \sum_{j=1}^N \mathbf{x}_j \end{aligned} \quad (\text{IV-37})$$

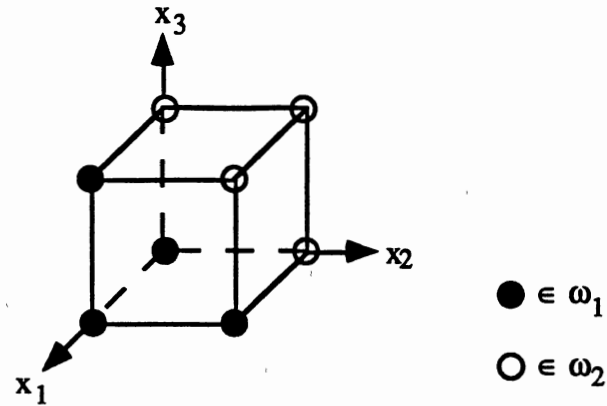


Figure IV-6. Original Observation Vectors

If the mean is not considered a random variable, the covariance may be similarly estimated using a sample average:

$$\begin{aligned}
 \mathbf{C} &= E[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^T] \\
 &= E[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x}^T - \boldsymbol{\mu}^T)] \\
 &= E[\mathbf{x}\mathbf{x}^T - \mathbf{x}\boldsymbol{\mu}^T - \boldsymbol{\mu}\mathbf{x}^T + \boldsymbol{\mu}\boldsymbol{\mu}^T] \\
 &= E[\mathbf{x}\mathbf{x}^T - 2\mathbf{x}\boldsymbol{\mu}^T + \boldsymbol{\mu}\boldsymbol{\mu}^T] \\
 &= E[\mathbf{x}\mathbf{x}^T] - 2E[\mathbf{x}\boldsymbol{\mu}^T] + E[\boldsymbol{\mu}\boldsymbol{\mu}^T] \\
 &= E[\mathbf{x}\mathbf{x}^T] - 2\boldsymbol{\mu}\boldsymbol{\mu}^T + \boldsymbol{\mu}\boldsymbol{\mu}^T \\
 &= E[\mathbf{x}\mathbf{x}^T] - \boldsymbol{\mu}\boldsymbol{\mu}^T \\
 &\approx -\boldsymbol{\mu}\boldsymbol{\mu}^T + \frac{1}{N} \sum_{j=1}^N \mathbf{x}_j \mathbf{x}_j^T
 \end{aligned} \tag{IV-38}$$

For each class, plugging in the observation vectors, we find that the means are unequal and the covariances are equal:

$$\mu_1 = \frac{1}{4} \begin{bmatrix} 3 \\ 1 \\ 1 \end{bmatrix} \quad \mu_2 = \frac{1}{4} \begin{bmatrix} 1 \\ 3 \\ 3 \end{bmatrix} \quad C = C_1 = C_2 = \frac{1}{16} \begin{bmatrix} 3 & 1 & 1 \\ 1 & 3 & -1 \\ 1 & -1 & 3 \end{bmatrix} \quad (\text{IV-39})$$

$$\delta = \mu_1 - \mu_2 = \frac{1}{4} \begin{bmatrix} 2 \\ -2 \\ -2 \end{bmatrix} \quad C^{-1} = \begin{bmatrix} 8 & -4 & -4 \\ -4 & 8 & 4 \\ -4 & 4 & 8 \end{bmatrix} \quad (\text{IV-40})$$

To maximize divergence in this special case, choose the transformation matrix as the transpose of the only nonzero eigenvalue's normal eigenvector of $C^{-1}\delta\delta^T$ (Tou and Heydorn 1967):

$$C^{-1}\delta\delta^T = \frac{1}{4} \begin{bmatrix} 1 & -1 & -1 \\ -1 & 1 & 1 \\ -1 & 1 & 1 \end{bmatrix} \quad (\text{IV-41})$$

$$\lambda = \frac{3}{4} \quad \mathbf{e} = \begin{bmatrix} -1 \\ 1 \\ 1 \end{bmatrix} \quad (\text{IV-42})$$

$$\mathbf{A} = \mathbf{e}^T = [-1 \quad 1 \quad 1] \quad (\text{IV-43})$$

$$\mathbf{y} = \mathbf{A}\mathbf{x} \quad (\text{IV-44})$$

$$\begin{array}{cccc} y_{11} = 0 & y_{12} = -1 & y_{13} = 0 & y_{14} = 0 \\ y_{21} = 1 & y_{22} = 1 & y_{23} = 2 & y_{24} = 1 \end{array} \quad (\text{IV-45})$$

A perfect discrimination rule would be to choose class 2 if the image pattern is greater than zero. These image patterns are nonoverlapping between the classes and, hence, the

3-D observation vectors have been successfully mapped to 1-D points with perfect discrimination. For comparison, the KLE transformation to 1-D fails to perfectly discriminate the data (Tou and Gonzalez 1974).

Bhattacharyya Distance

The calculation of error probability is a difficult task, even when the observation vectors have a normal pdf. Closed-form expressions for probability of error exist only for trivial, uninteresting situations. Often the best we can hope for is a closed-form expression of some upper bound of error probability. The Bhattacharyya distance is closely tied to the probability of error as an upper bound on the Bayes error for normally distributed classes (Fukunaga 1990). For normal pdf's, the Bhattacharyya distance between class ω_1 and ω_2 , also referred to as $\mu(\frac{1}{2})$, is:

$$d_B^2 = \frac{1}{2} \ln \frac{\frac{|C_i + C_j|}{2}}{|C_i|^{1/2} |C_j|^{1/2}} + \frac{1}{2} (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^T \left(\frac{C_i + C_j}{2} \right)^{-1} (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j) \quad (\text{IV-46})$$

The Bhattacharyya distance directly compares the estimated mean vector and covariance matrix of the test segment with those of the target speaker. If inclusion of the test covariance in the metric is useful, Bhattacharyya distance will outperform Mahalanobis distance. Neglecting scaling, the second term is the Mahalanobis distance using an average covariance matrix. As will be shown later, if the Mahalanobis distance using an average covariance matrix performs poorly, a different pair of scale factors can yield better discrimination.

Note that Equation IV-46 is the sum of two components, one is based solely upon the covariance matrices and the other involves differences between the mean vectors. These components can be characterized respectively as an average shape and the difference in

size of the pdf's. This shape component, the *Bhattacharyya shape*, will prove very useful later on:

$$d'_B = \ln \frac{\frac{|C_i + C_j|}{2}}{|C_i|^{1/2} |C_j|^{1/2}} \quad (\text{IV-47})$$

The Bhattacharyya distance and the divergence measure have many similarities (Devijver 1974; Kailath 1967; Lee 1991). As will be seen later, they both yield similar speaker identification performance.

The next chapter introduces statistical pattern matching and receiver operating curves.

CHAPTER V

PATTERN MATCHING

The pattern matching task of speaker verification involves computing a match score, which is a measure of the similarity of the input feature vectors to some model. Speaker models are constructed from the features extracted from the speech signal. To enroll users into the system, a model of the voice, based on the extracted features, is generated and stored (possibly on an encrypted smart card). Then, to authenticate a user, the matching algorithm compares/scores the incoming speech signal with the model of the claimed user.

There are two types of models: template models and stochastic models. In stochastic models, the pattern matching algorithm is probabilistic, typically a likelihood measure. For template models, the pattern matching algorithm is distance based. Likelihood measures can be approximated in template based models by scoring against the claimed speaker model versus a global speaker model (Higgins 1990).

The template model and its corresponding distance measure is perhaps the most intuitive method. The template method can be dependent or independent of time. An example of a time-independent template model is vector quantization modeling (Soong and others 1987). All temporal variation is ignored in this model and global averages (e.g., centroids) are all that is used. A time-dependent template model is more complicated because of human speaking rate variability.

Template Models

The simplest template model consists of a single template, \bar{x} , which is the model for a frame of speech. The match score between the template, \bar{x} , for the claimed speaker and an input feature vector, x_i , from the unknown user is given by $d(x_i, \bar{x})$. The model for the claimed speaker could be the centroid (mean) of a set of N training vectors:

$$\bar{x} = \mu = \frac{1}{N} \sum_{i=1}^N x_i \quad (\text{V-1})$$

Many different distance measures between the vectors x_i and \bar{x} can be expressed as:

$$d(x_i, \bar{x}) = (x_i - \bar{x})^T W (x_i - \bar{x}) \quad (\text{V-2})$$

where W is a weighting matrix. If W is an identity matrix, the distance is *Euclidean*; if W is the inverse covariance matrix corresponding to mean \bar{x} , then this is the *Mahalanobis distance*, as shown in Equation IV-4. The Mahalanobis distance gives less weight to the components having more variance and is equivalent to a Euclidean distance on principal components, which are the eigenvectors of the original space as determined from the covariance matrix (Duda and Hart 1973).

Dynamic Time Warping

The most popular method to compensate for speaking rate variability is known as dynamic time warping (DTW) (Sakoe and Chiba 1978). A text-dependent template model is a sequence of templates $(\bar{x}_1, \dots, \bar{x}_N)$ which must be matched to an input sequence (x_1, \dots, x_M) . In general, N is not equal to M because of timing inconsistencies in human speech. The asymmetric match score, z , is given by:

$$z = \sum_{i=1}^M d(x_i, \bar{x}_i) \quad (\text{V-3})$$

where the template index, j , is typically given by a dynamic time warping algorithm. Given reference and input signals, the DTW algorithm does a constrained, piecewise linear mapping of one (or both) time axis(es) based on a minimum distance criteria to align the two signals. At the end of the time warping, the accumulated distance is the basis of the match score. Instead of using global averages, this method accounts for the normalized variation over time (trajectories) of parameters. This corresponds to the dynamic configuration of the human articulators and vocal tract. For example, Figure V-1 shows what a warp path might look like if the energies between two speech signals are used as warp features.

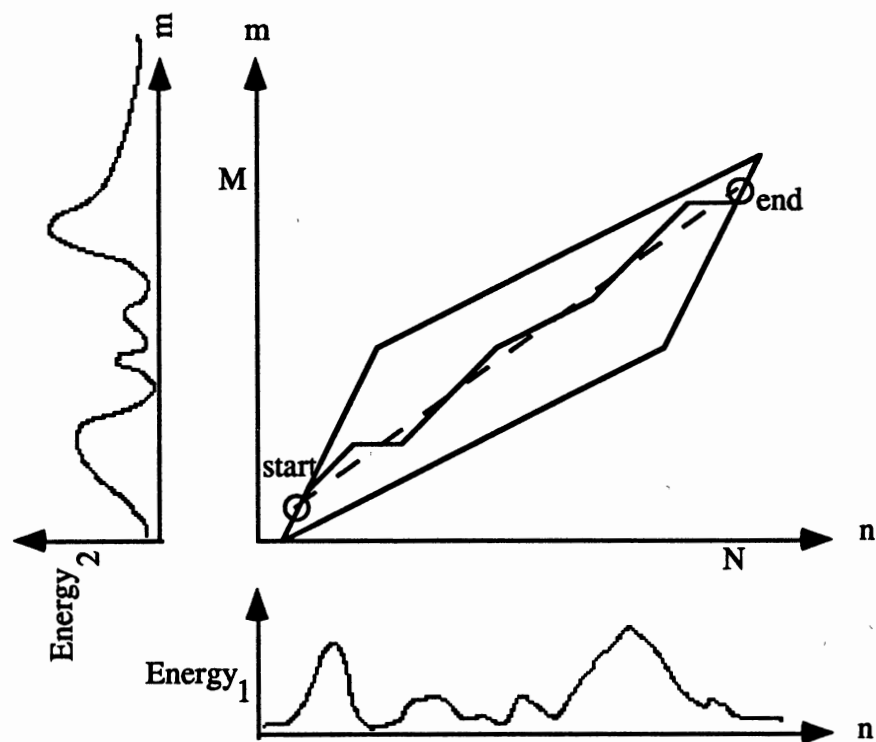


Figure V-1. Dynamic Time Warping Two Energy Signals

If the warp signals were identical, the warp path would be a diagonal line and the warping would have no effect. The Euclidean distance between the two signals in the energy domain is the accumulated deviation off the dashed diagonal warp path. The parallelogram surrounding the warp path represents the Sakoe slope constraints of the warp, which act as boundary conditions to prevent excessive warping over a given segment.

Vector Quantization Source Modeling

Another form of template model uses multiple templates to represent a frame of speech and is referred to as vector quantization source modeling (Soong and others 1987). A VQ code book is designed by standard clustering procedures for each enrolled speaker using his training data based upon reading a specific text. The pattern match score is the distance between an input vector and the minimum distance codeword in the VQ code book C . The match score is:

$$z = \sum_{j=1}^L \min_{\bar{x} \in C} \{d(\mathbf{x}_j, \bar{x})\} \quad (\text{V-4})$$

The clustering procedure, used to form the code book, averages out temporal information from the codewords. Thus, there is no need to perform a time alignment. The lack of time warping greatly simplifies the system; however, there is likely to be some speaker dependent information that is lost. The disadvantage of this approach is that it ignores temporal information.

Nearest Neighbors

A new method combining the strengths of the DTW and VQ methods is called nearest neighbors (NN) (Higgins 1990). Unlike the VQ method, the NN method does not cluster

the enrollment training data to form a compact code book. Instead, it keeps all the training data and can, therefore, use temporal information.

As shown in Figure V-2, the interframe distance matrix is computed by measuring the distance between test session frames (the input) and the claimant's enrollment session frames (stored). The nearest neighbor distance is the minimum distance between a test session frame and the enrollment frames. The nearest neighbor distances for all the test session frames are then averaged to form a match score. Similarly, as shown in the rear planes of Figure V-2, the test session frames are also measured against a set of stored reference speakers to form match scores. The match scores are then combined to form a likelihood approximation.

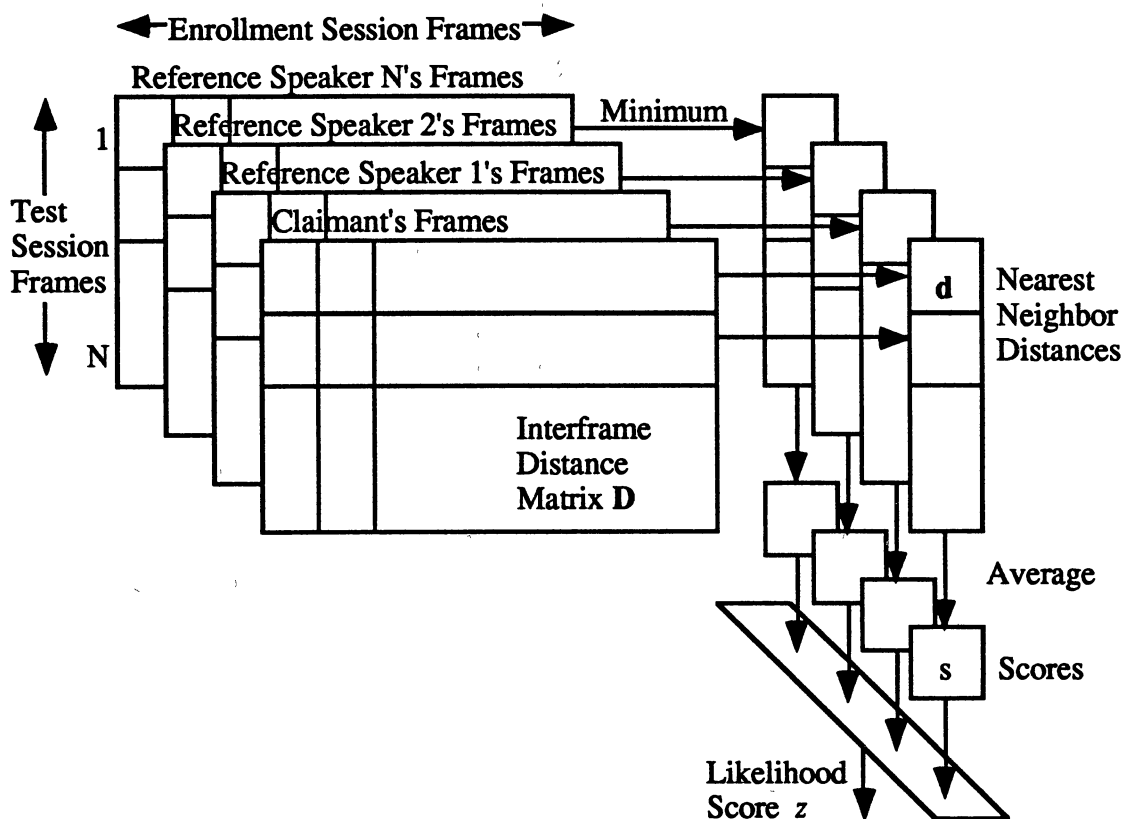


Figure V-2. Nearest Neighbor Method

The NN method is one of the most memory and compute intensive speaker authentication algorithms. It is also one of the most powerful methods, as illustrated later in Figure VII-1.

Stochastic Models

Template models have dominated work in text-dependent speaker recognition. The distance measure approach is an intuitively reasonable test of similarity, but stochastic models can offer more flexibility and result in a more theoretically meaningful score.

Using a stochastic model, the pattern matching problem can be formulated as measuring the likelihood of an observation (a feature vector of a collection of vectors from the unknown speaker) given the speaker model. The observation is a random vector with a conditional pdf which depends upon the model corresponding to the class of the observation (the claimed identity or an impostor). The conditional pdf of the feature vector can be estimated from a set of training vectors and, given the estimated density, the probability that the observation was generated by the claimed speaker can be determined.

The estimated pdf can either be a parametric or nonparametric model. From this model, for each frame of speech (or average of a sequence of frames), the probability that it was generated by the claimed speaker can be estimated. This probability is the match score. If nothing is known about the true densities, then nonparametric statistics can be used to find the match score. If the model is parametric, then a specific pdf is assumed and the appropriate parameters of the density can be estimated using the maximum likelihood estimate. For example, one useful parametric model is the multivariate normal model. Unbiased estimates for the parameters of this model, the mean μ and the covariance C , are given by Equations IV-6 and IV-7, respectively. In this case, the probability that an observed feature vector, x , was generated by the model is:

$$p(\mathbf{x}_i|\text{model}) = (2\pi)^{-k/2} |\mathbf{C}|^{-1/2} \exp\left\{-\frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu})^T \mathbf{C}^{-1}(\mathbf{x}_i - \boldsymbol{\mu})\right\} \quad (\text{V-5})$$

Hence, $p(\mathbf{x}_i|\text{model})$ is the match score.

The match scores for text-dependent models are given by the probability of a sequence of frames without assuming independence of speech frames. The model represents a specific sequence of speech frames. One stochastic model that is very popular in modeling sequences is the hidden Markov model (HMM). The HMM is a finite-state machine, where each state, s_i , is associated with a pdf (or feature vector stochastic model), $p(\mathbf{x} | s_i)$. The states are connected by a transition network, where the state transition probabilities are $a_y = p(s_j | s_i)$. For example, a hypothetical three-state HMM is illustrated in Figure V-3.

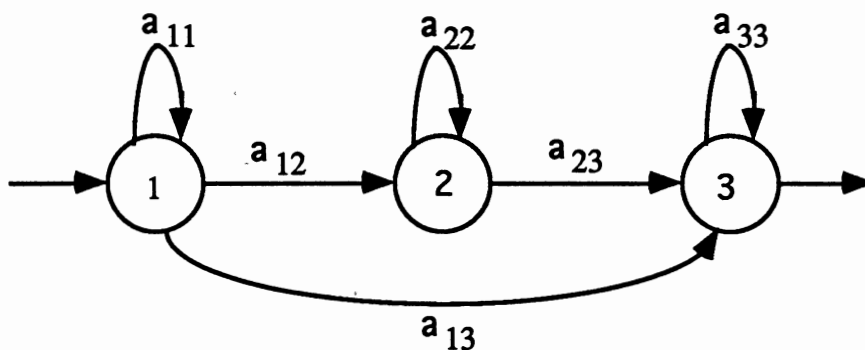


Figure V-3. An Example of a Three-State Hidden Markov Model

The probability that a sequence of speech frames was generated by this model is found by using Baum-Welch decoding (Rabiner and Juang 1986; Rabiner 1989). This probability is the score of the input speech given the model:

$$p(\mathbf{x}(1;L) | \text{model}) = \sum_{\substack{\text{all state} \\ \text{sequences}}} \prod_{i=1}^L p(\mathbf{x}_i | s_i) p(s_i | s_{i-1}) \quad (\text{V-6})$$

This might be a more theoretically meaningful score. HMMs were not further pursued in this research because the recent results of Tishby showed this method to be comparable in performance to conventional VQ methods (Tishby 1991).

Classification methods and statistical decision theory is presented in the following chapter.

CHAPTER VI

CLASSIFICATION AND DECISION THEORY

Having computed a match score between the input speech feature vector and a model of the claimed speaker's voice, a decision is made whether to accept or reject the speaker or ask for another token. The accept or reject decision process is actually an accept, continue, time-out, or reject hypothesis-testing problem. Thus, the decision making, or classification, procedure is a sequential hypothesis-testing problem (Wald 1947).

Hypothesis Testing

Given a match score, the classification problem involves choosing between two hypotheses: that the user is the claimed speaker or that he is an impostor. Let H_0 be the hypothesis that the user is an impostor and let H_1 be the hypothesis that the user is, indeed, the claimed speaker. As shown in Figure VI-1, the match scores of the observations form two different pdf's according to whether the user is the claimed speaker or an impostor.

The names of the probability areas (or volumes in the case of multidimensional match scores) in Figure VI-1 are given in Table VI-1. To find a given performance probability volume, the hypothesis determines over which pdf to integrate and the threshold determines which decision region forms the limits of integration.

Let $p(z|H_0)$ be the conditional density function of the observation score, z , generated by an impostor and likewise $p(z|H_1)$ for the claimed speaker. If the true conditional score densities for the claimed speaker and the impostor are known, then the Bayes test

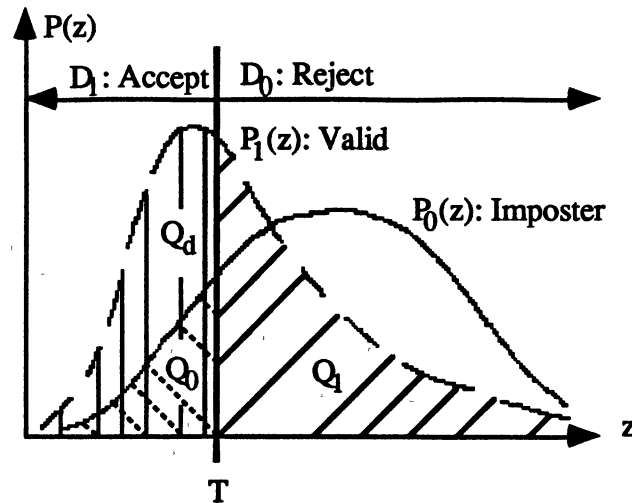


Figure VI-1. Valid and Impostor Densities

for minimum error, with equal misclassification costs, for speaker A is based upon the likelihood ratio for speaker A, $\lambda_A(z)$ (Fukunaga 1990):

$$\lambda_A(z) \equiv \frac{p_A(z|H_0)}{p_A(z|H_1)} \quad (\text{VI-1})$$

Figure VI-2 shows an example of two score pdf's. The probability of error is determined by the amount of overlap in the two pdf's. The smaller the overlap between the two pdf's, the smaller the probability of error. The overlap in two Gaussian pdf's with means μ_0 and μ_1 and equal variance σ can be measured by the F-ratio:

$$F = \frac{(\mu_0 - \mu_1)^2}{\sigma^2} \quad (\text{VI-2})$$

If the true conditional score densities for the claimed speaker and the impostor are unknown, the two pdf's can be estimated from sample experimental outcomes. The conditional pdf given true speaker A, $p_A(z|H_1)$, is estimated from the speaker's own

TABLE VI-1
PROBABILITY TERMS AND DEFINITIONS

Performance Probabilities	D	H	Name of Probability	Decision Result	
				Type I error	False acceptance or alarm
Q_0	1	0	Size of test "significance"	Type I error	False acceptance or alarm
Q_1	0	1		Type II error	False rejection
$Q_d = 1 - Q_1$	1	1	Power of test		True acceptance
$1 - Q_0$	0	0			True rejection

scores using his model. The conditional pdf given an impostor, $p_A(z|H_0)$, is estimated from other speakers' scores using speaker A's model.

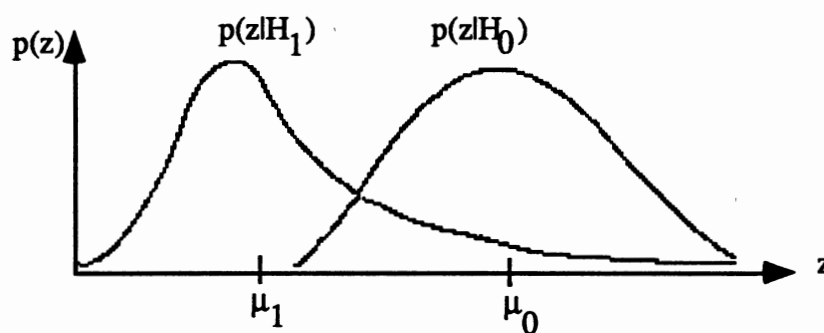


Figure VI-2. An Example of Score Densities

Now that the likelihood ratio for speaker A , $\lambda_A(z)$, can be determined, the classification problem can be stated as choosing a threshold, T , so that the decision rule is:

$$\text{if } \lambda_A(z) \begin{cases} \geq T, \text{ choose } H_0 \\ < T, \text{ choose } H_1 \end{cases} \quad (\text{VI-3})$$

The threshold, T , can be determined by: (1) setting T equal to an estimate of p_1/p_0 to approximate minimum error performance, where p_0 and p_1 are the a priori probabilities that the user is an impostor and that the user is the true speaker, respectively; (2) choosing T to satisfy a fixed *false acceptance* (FA) or *false rejection* (FR) criterion (Neyman-Pearson); or (3) varying T to find different FA/FR ratios and choosing T to give the desired FA/FR ratio. With cautious constraints, T could be made speaker specific, speaker adaptive, and/or risk adaptive (e.g., break-ins may be more likely at night).

Receiver Operating Curve

Since either of the two types of errors can be reduced at the expense of an increase in the other, a measure of overall system performance must specify the levels of both types of errors. The tradeoff between FA and FR is a function of the decision threshold. This is depicted in the receiver operating curve (ROC), which plots probability of FA versus probability of FR (or FA rate versus FR rate). For example, Figure VI-3 shows a hypothetical family of ROCs plotted on a log-log scale. The line of equal error probability is shown as a dotted diagonal line. The family of lines at -45 degrees represents systems with different FA•FR products, with better systems being closer to the origin. For any particular system, the ROC is traversed by changing the threshold of acceptance for the likelihood ratio. The straight line ROCs in Figure VI-3 indicate that the product of the probability of FA and the probability of FR is a constant for this hypothetical system (this is not true in general) and is equal to the square of what is

referred to as the equal error rate (EER). The EER is the value for which the type I errors and type II errors are equal.

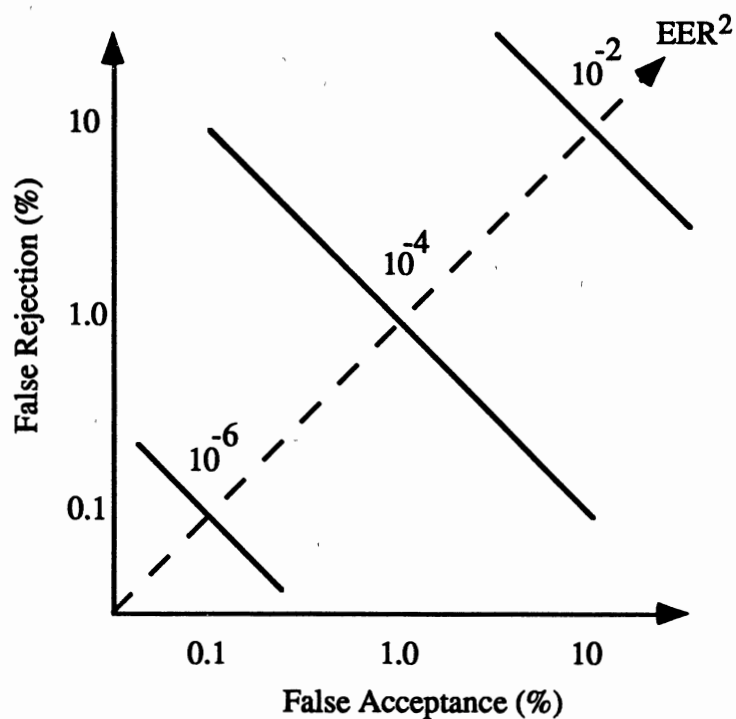


Figure VI-3. Hypothetical Receiver Operating Curves

Data Fusion

Combining fundamentally different features is the topic of data fusion (Hedges and Olkin 1985). In speaker authentication, features arise from physiologically different phenomena. For example, some features may correspond to learned traits and others to physical ones. They may also have different sampling rates. Different test methods might

be useful in reducing the error rate. The tools of data fusion may provide a mathematical foundation for combining these disparate features and different tests.

For example, the results show different wolf and sheep populations for two different authentication systems. Data fusion methods could allow these systems to be merged into a single system whose performance is more powerful than either system, alone.

The next chapter demonstrates the speaker identification performance of the new algorithm relative to two reference speaker verification algorithms.

CHAPTER VII

PERFORMANCE

Using the YOHO prerecorded speaker authentication database, the following results on wolves and sheep were measured. The impostor testing was simulated by randomly selecting a valid user (a potential wolf) and altering their identity claim to match that of a randomly selected target user (a potential sheep). Because the potential wolf is not intentionally attempting to masquerade as the potential sheep, this is referred to as the “casual impostor” paradigm. The YOHO database has 10 test sessions for each of 186 subjects. For only 1 test session, there are $186C_2 = 17,205$ pairwise combinations. Because of computational limitations, it is impractical to test all pairwise combinations for all 10 test sessions. Thus, the simulated impostor testing randomly drew across the 10 test sessions. Testing the system to a certain confidence level implies a minimum requirement for the number of trials. In this testing, there were 9300 simulated impostor trials to test to the desired confidence (Higgins 1990).

DTW System

Table VII-1 shows two measures of wolves and sheep for the DTW system: those who were wolves or sheep at least once and those who were wolves or sheep at least twice. Thus, type I errors are spread across a very narrow portion of the population, especially if two errors are required to designate a person as a wolf or sheep. The difficulty in acquiring enough data to adequately represent the wolf and sheep problem is perhaps the main reason why there has been relatively little work specifically directed at understanding and improving type I errors.

TABLE VII-1
KNOWN WOLVES AND SHEEP, DTW SYSTEM

186 Subjects of the YOHO Database	
At least one Type I Error	At least two Type I Errors
17 Wolves (9%)	2 Wolves (1%)
11 Sheep (6%)	5 Sheep (3%)

From the 9300 trials, there were 19 type I errors for the DTW system. Table VII-2 shows that these 19 pairs of wolves and sheep have interesting sexual relationships. Even though the database contains four times as many males as it does females, the ratio of male wolves to female wolves (18:1) seems disproportionate. It's also interesting to note that one male wolf successfully preyed upon three different female sheep.

The YOHO database provides at least 19 pairs of wolves and sheep under the DTW system for further investigation. It should be noted that because of computational limitations, not all possible wolf and sheep combinations have been tested. Even with this massive database, relatively few wolves and sheep have been discovered to date.

ROC of DTW and NN Systems

The ROC in Figure VII-1 was made on the nearest neighbor system using the YOHO database. The log-log plot has the same axes scaling so it is easy to see the 0.5% equal error rate on the dashed diagonal line. The NN system meets the U.S. Government's performance requirement of 0.1% FA and 1% FR. The NN system is the first one known to meet this level of performance.

TABLE VII-2
WOLF AND SHEEP SEXUAL RELATIONSHIPS

19 type I errors across 9300 impostor trials		
Number of type I errors	Wolf sex	Sheep sex
15	males	males
1	female	female
3	1 male	3 females

The U.S. Government's goal of 0.01% FA and 0.1% FR is an order of magnitude beyond the required performance by an order of magnitude in each dimension. Extraordinary improvements in the state of the art of speaker authentication will be required to meet this goal. Because of the more demanding false acceptance objective, my research focused on false acceptance errors.

Figure VII-1 shows the NN system's receiver operating curve and a point on the ROC of the DTW system; ROCs of better systems are toward the origin. The NN system outperforms the DTW system by about half an order of magnitude. More importantly, the NN system meets the U.S. Government's performance requirement.

These overall error rates do not show the individual wolf and sheep populations of the two systems. As shown in the following sections, the two systems commit different errors. Perhaps these systems could be fused to exploit their respective strengths (Hedges and Olkin 1985).

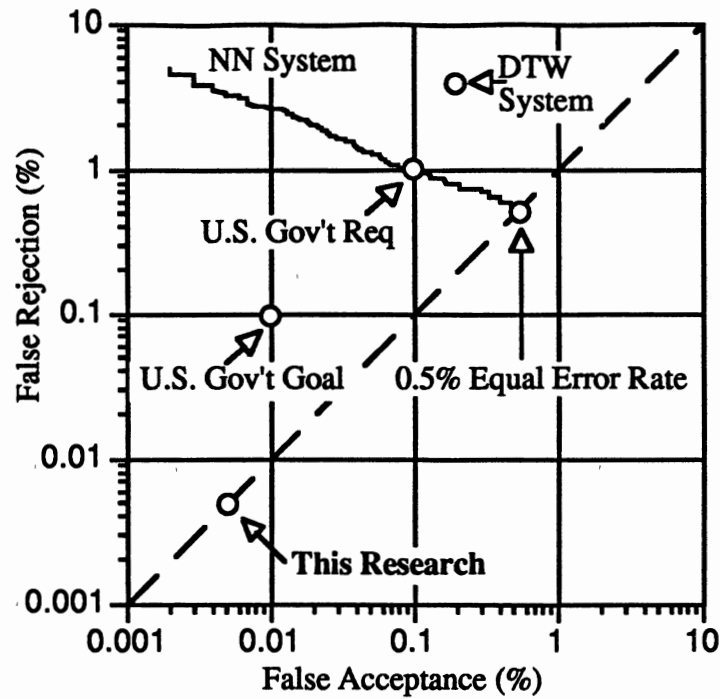


Figure VII-1. Receiver Operating Curves

Wolves and Sheep

Figure VII-2 shows the individual speakers who were falsely accepted as other speakers by the DTW system. The following 3-D histogram plots can be interpreted by example. In Figure VII-3, the person with an identification number of 97328 is a never a wolf and is a sheep once under the DTW system.

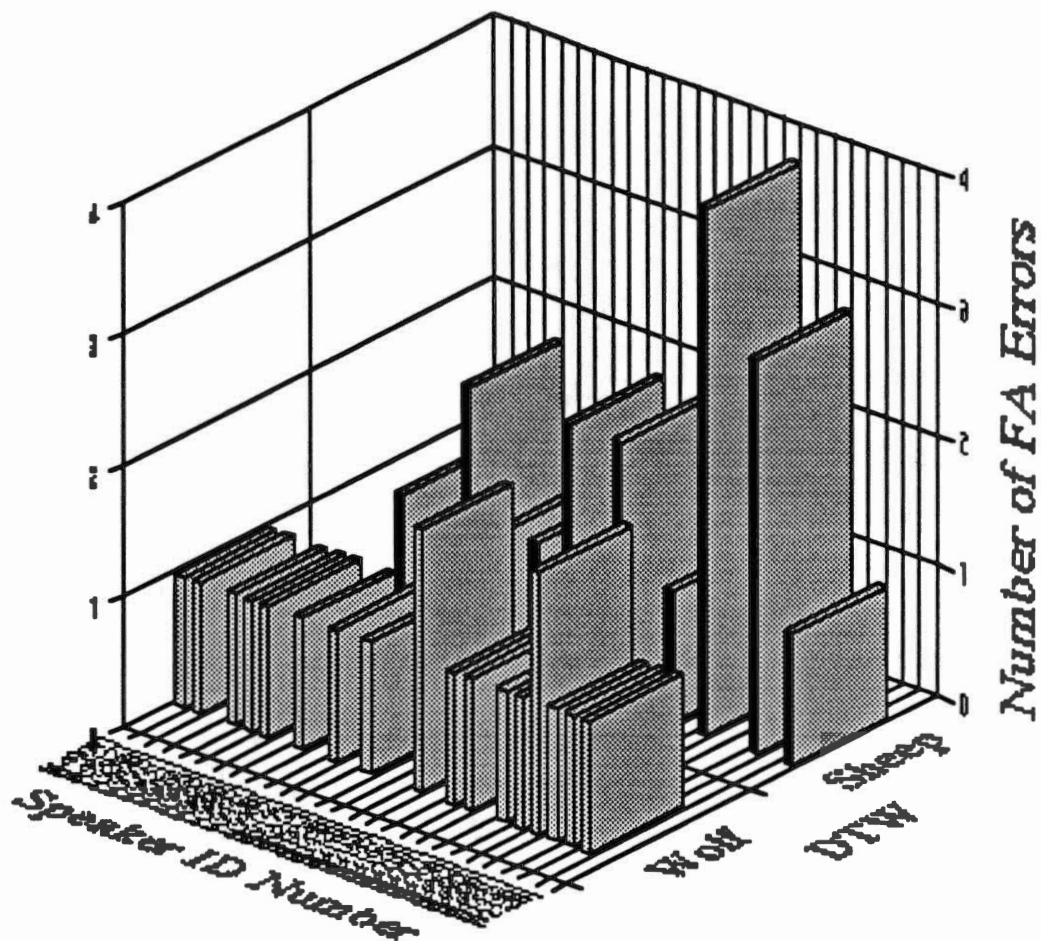


Figure VII-2. Speaker vs FA Errors for DTW System's Wolves and Sheep

To get a better angle on seeing if there are speakers who are both wolves and sheep, Figure VII-2 is rotated into Figure VII-3.

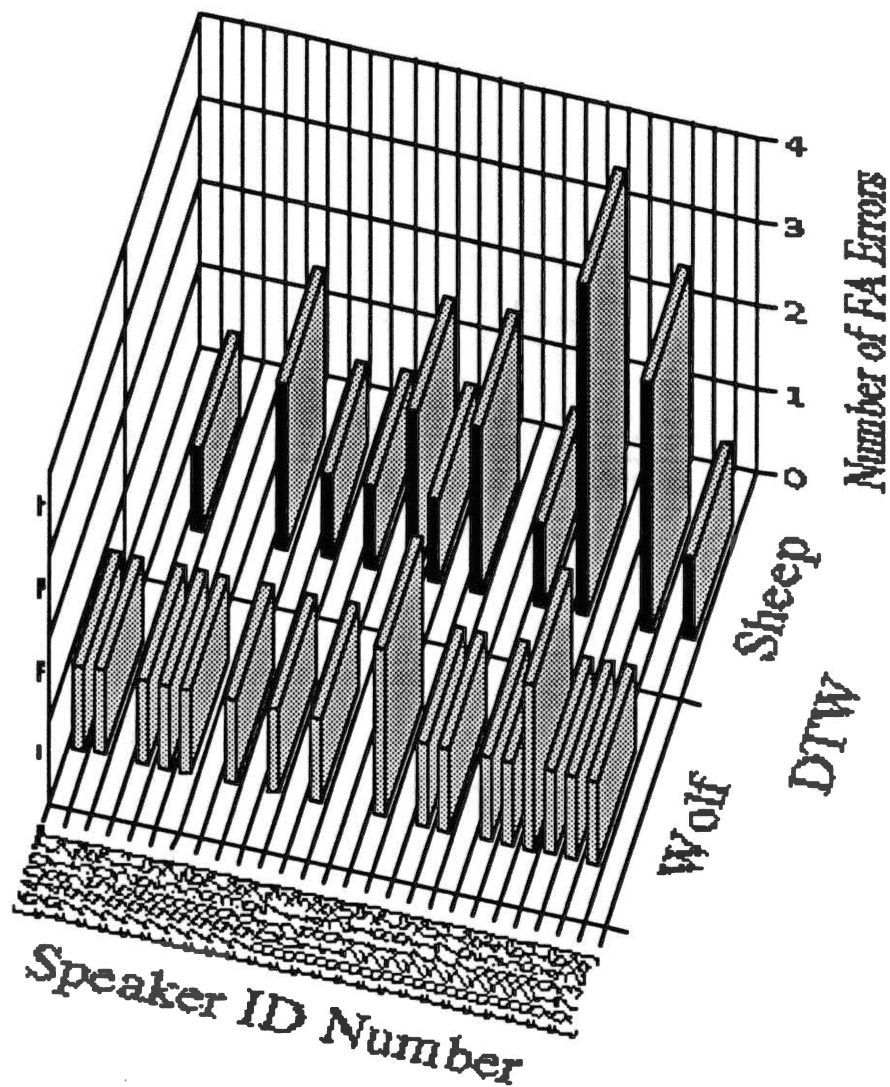


Figure VII-3. FA Errors for DTW System's Wolves and Sheep

The DTW system rarely has the same speaker as both a wolf and a sheep (only two exceptions in this data). These exceptions, called *wolf-sheep*, probably have poor models because they match a sheep's model more closely than their own and a wolf's model also matches their model more closely than their own. These *wolf-sheep* would likely benefit from retraining to improve their models.

Now, let's look at the Nearest Neighbor system. Figure VII-4 shows NN test sessions for which an impostor's training data matched the session better than the speaker's own training data.

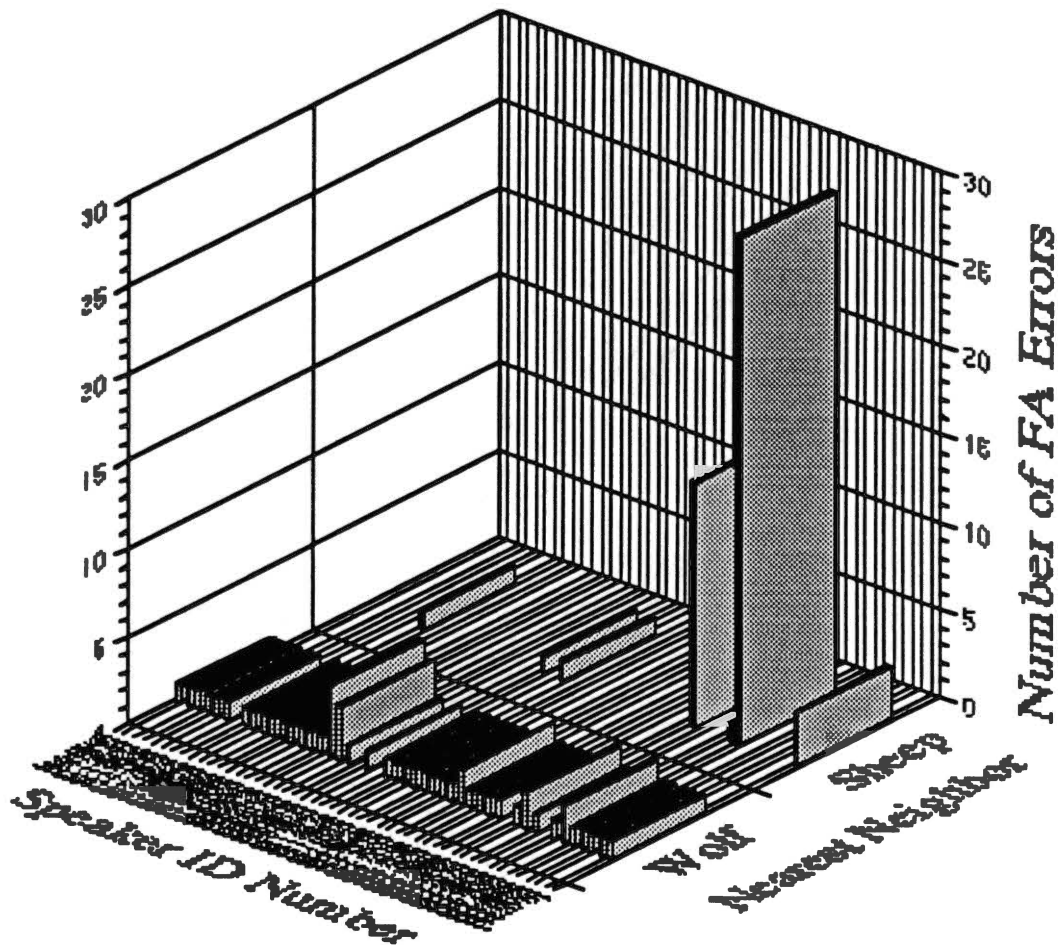


Figure VII-4. Speaker vs FA Errors for NN System's Wolves and Sheep

Two speakers, who are sheep, are seen to dominate the false acceptance errors. The NN system performance would be greatly improved if these two speakers were better handled by the system.

Now we'll investigate the relations between the NN and DTW systems. Figure VII-5 shows the sheep of the NN and DTW systems.

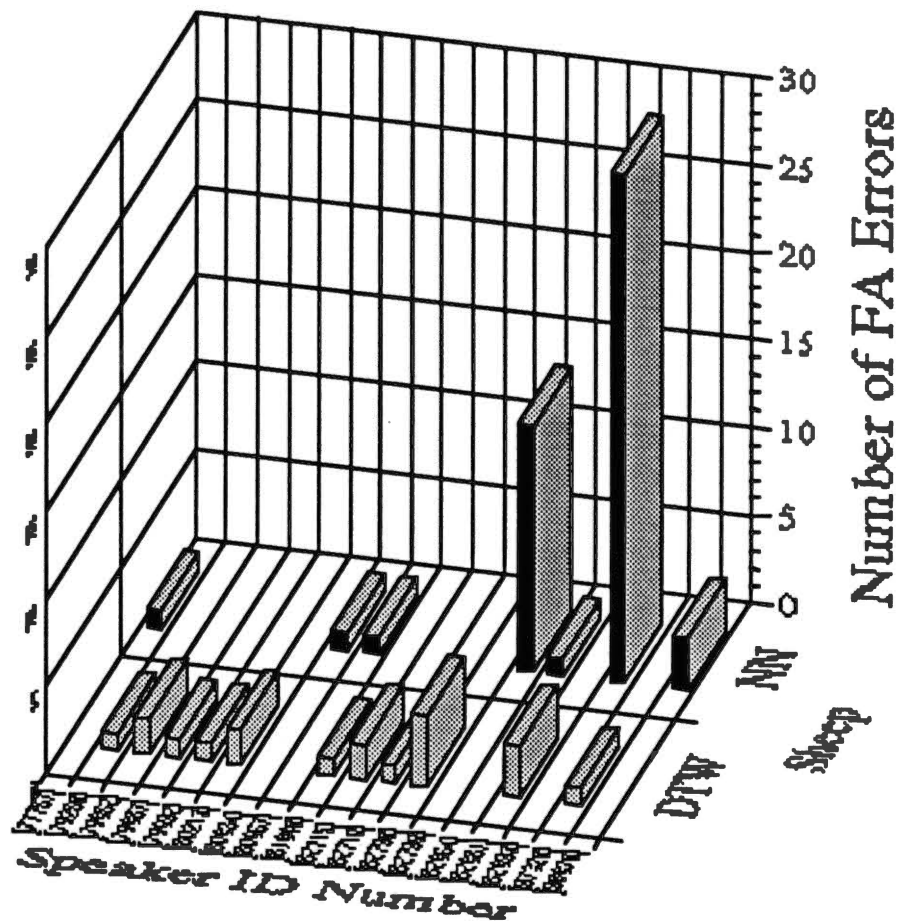


Figure VII-5. Speaker vs FA Errors for DTW and NN Systems' Sheep

It should be noted from Figure VII-5 that the two sheep who dominate the FA errors of the NN system were not found to be sheep in the DTW system. This suggests the potential for making a significant performance improvement by combining the systems.

Figure VII-6 shows that the wolves of the NN system are dominated by a few individuals who do not cause errors in the DTW system. Again, this suggests the

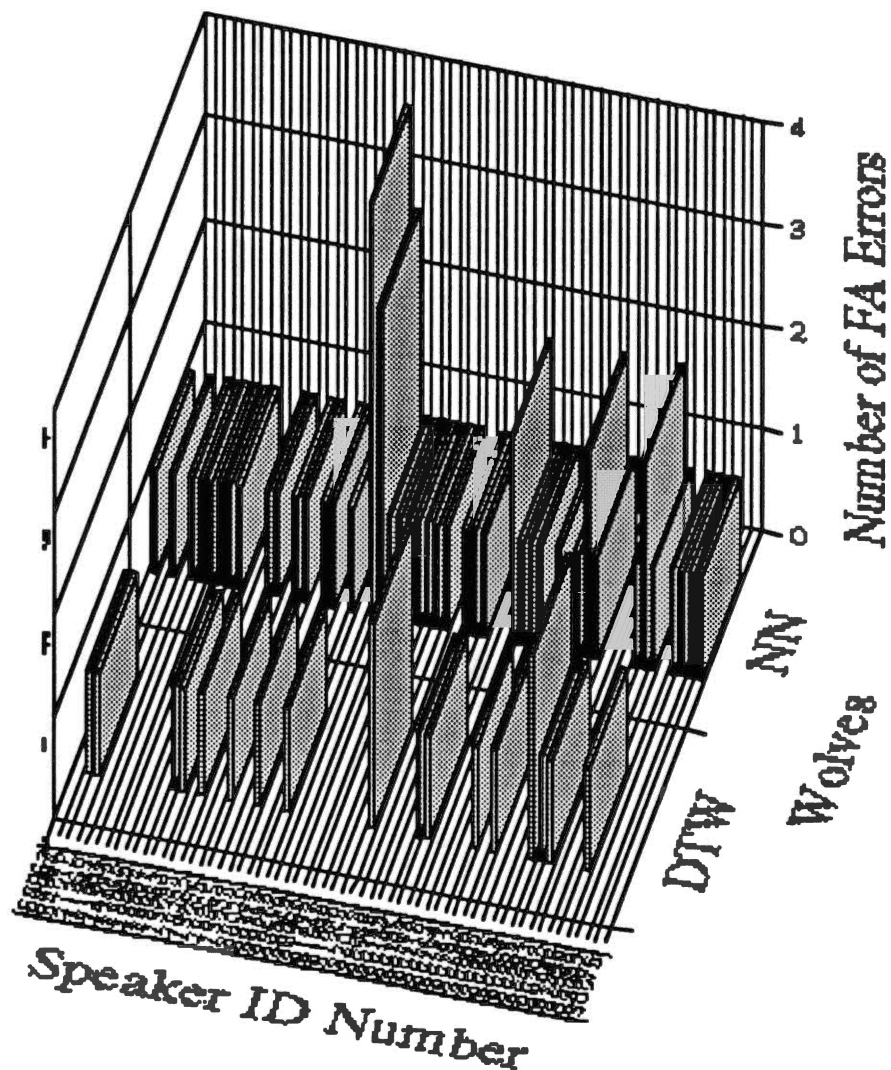


Figure VII-6. Speaker vs FA Errors for DTW and NN Systems' Wolves

potential for realizing a performance improvement by combining elements of the NN and DTW systems.

Figure VII-7 shows the number of false acceptance errors that occur for each test session of the NN system. The figure clearly shows that a couple sessions, namely numbers 880 and 1858, have an inordinate number of false acceptance errors. Something appears to be wrong with these sessions. Upon listening to sessions 880 and 1858, it sounds like these sessions have more boominess than the other test (and enrollment) sessions. It's possible that the acoustic environment changed in between the sessions.

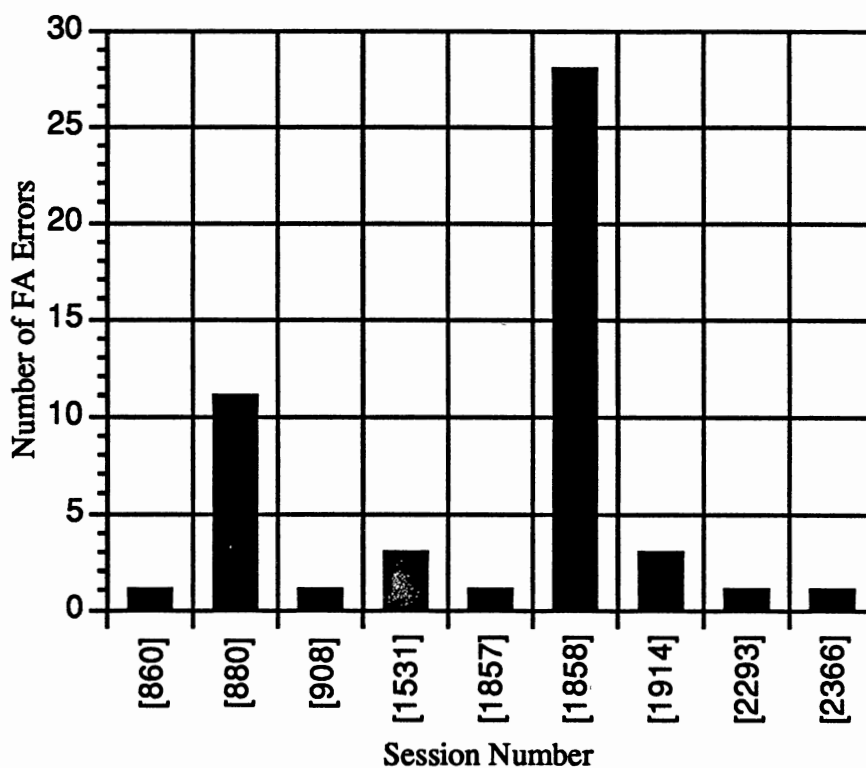


Figure VII-7. FA Errors vs Session Number for NN System

For example, an open door into a reflective room could add boominess to the sound and alter the spectral features.

Wolves and sheep come in pairs. Figure VII-8 shows the DTW system's wolf and sheep pairings for the YOHO database.

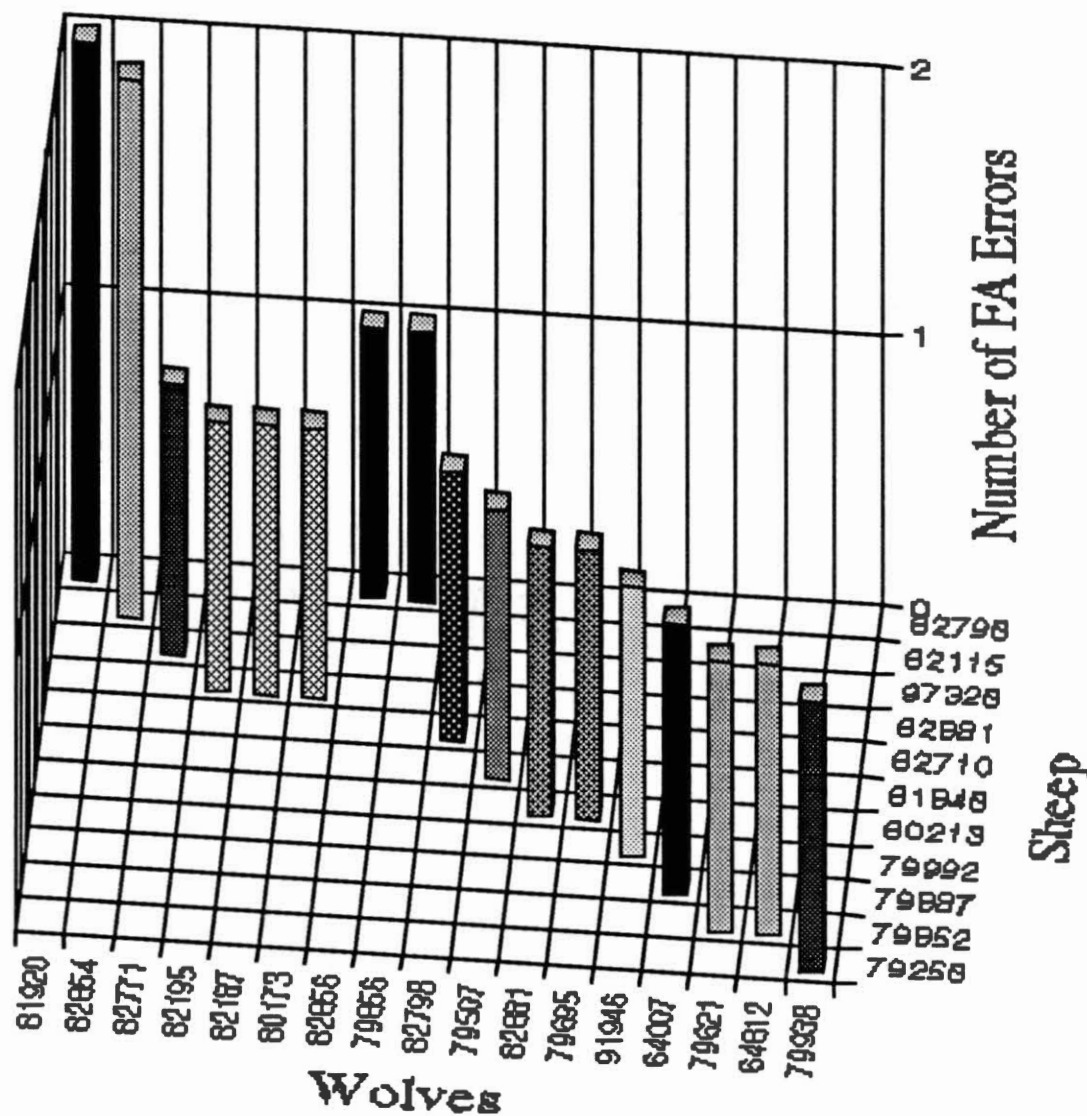


Figure VII-8. Wolf and Sheep Pairings of the DTW System

It should be noted that, under the DTW system, speaker 82798 is a particularly vulnerable sheep with respect to wolves 81920, 82866, and 79866. These speakers, in addition to the others shown in Figure VII-8, will be of prime interest in the following experiments.

LSP Divergence Shape Speaker Identification

A speaker identification test using motivated speakers, a high-quality stationary channel, and constrained grammar yielded 99.95% correct speaker identification. This experiment uses 44 people from the YOHO database with 80 seconds of speech for training and a separate 80 seconds of speech for testing. Each speaker is compared to a different session of himself and to 2 sessions of 43 other speakers. The “closest” speaker to each candidate is identified. In one experiment, only 1 false identification error was made on a total of 1936 tests. The line spectrum pair frequency features measured by the divergence shape (an information-theoretic divergence measure without mean information) “closeness” criterion yielded this result. This outperformed the LSP Bhattacharyya shape (2 errors), the LSP Bhattacharyya distance (4 errors), and the LSP divergence measure (3 errors).

In the following mesh plots, each of the 44 people are shown along the x- and y-axes; the x-axis represents speech collected from session 1 versus the y-axis with speech collected from session 2. Thus, there are 44^2 tests, each represented by a point on the mesh. The z-axis is the reciprocal of the measure indicated in the figure’s caption using LSP features. Thus, “close” speakers will cause a peak in the z-axis. The ideal structure, representing perfect speaker identification, would be a prominent diagonal such that $a_{ii} > a_{ij} \forall i \neq j$.

Notice the nearly ideal prominent diagonal structure in Figure VII-9 provided by the LSP divergence shape; thus, its discrimination power is very strong. The single error made by the LSP divergence shape, shown by an arrow in Figure VII-9, is between session 1 of speaker 59771 and session 2 of speaker 79082. It’s interesting to note that

this is not one of the DTW system's pairs of wolves and sheep as shown in Figure VII-8. It's also interesting to note that this same error occurs in all the LSP based divergence and Bhattacharyya distance systems as shown by a peak in the following mesh plots at the same location as the arrow in Figure VII-9.

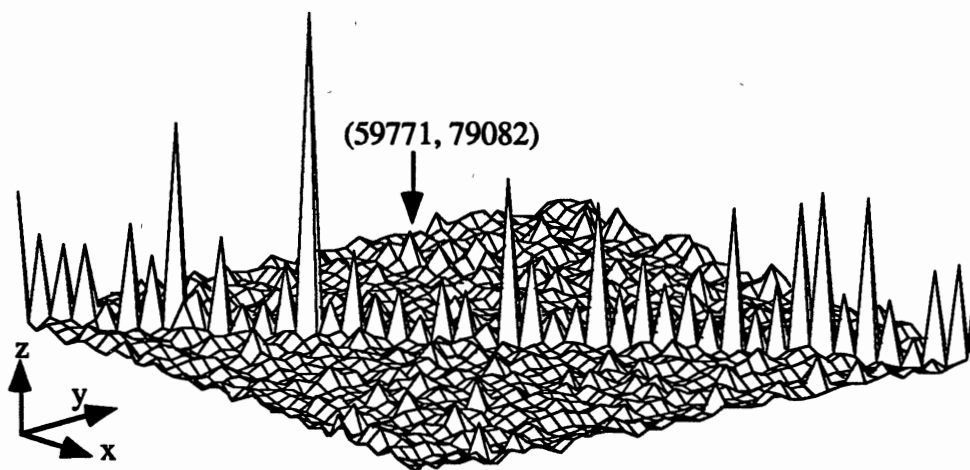


Figure VII-9. LSP Divergence Shape (1 error)

Notice the similarity in structure between the mesh plots of the LSP Bhattacharyya shape shown in Figure VII-10 and the LSP divergence shape. Not only do these measures perform similarly well, but the measures also appear to be closely related.

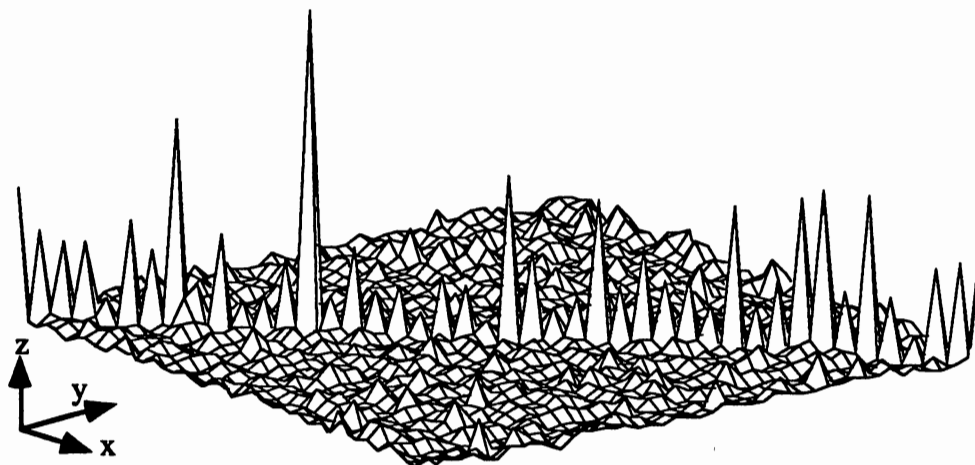


Figure VII-10. LSP Bhattacharyya Shape (2 errors)

Note the degraded performance of the LSP Bhattacharyya distance, Figure VII-11, versus the LSP Bhattacharyya shape. Including the means in the Bhattacharyya distance degraded its performance. This discovery provided the insight toward the development of the shape measures.

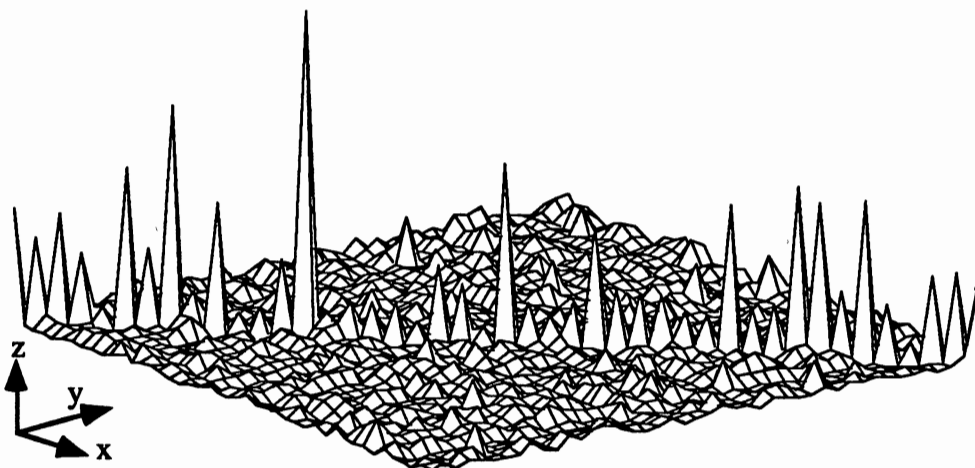


Figure VII-11. LSP Bhattacharyya Distance (4 errors)

Note the degraded performance of the LSP divergence measure, Figure VII-12, relative to the divergence shape. Again, including the means degraded performance.

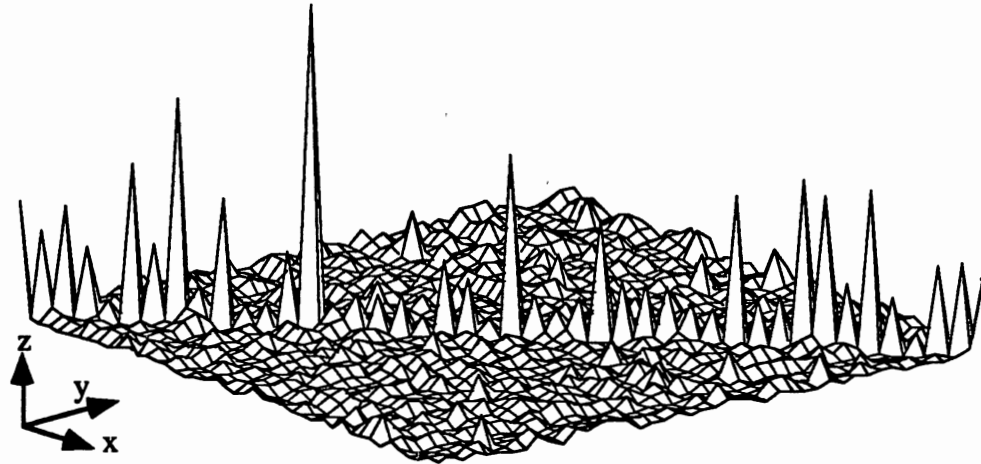


Figure VII-12. LSP Divergence Measure (3 errors)

The power of using the LSP features in these measures is clearly shown by the prominent diagonal structure in the previous figures.

The author's previous results are summarized in Table VII-3, with additional identification experiments performed on the same data. Out of the 1936 tests, Euclidean distance commits 38 errors (1.96% error) and Mahalanobis distance makes 21 errors (1.08%) using LP cepstrum combined with LAR features. The LSP divergence shape performs the best among these experiments with only 1 error (0.05%).

TABLE VII-3
ERRORS OF VARIOUS FEATURES AND MEASURES

	LSP	Cep	LAR
Divergence Shape	0.05%	0.15%	
Bhattacharyya Shape	0.10%	0.10%	
Bhattacharyya Distance	0.21%	0.10%	
Divergence Measure	0.15%	0.21%	0.52%
Mahalanobis Distance		1.08%	
Euclidean Distance		1.96%	

One might conclude from these results that the means of the features tested tend to be unreliable, while the variances and covariances in the features have strong discrimination power. In fact, the author was led to the divergence shape and Bhattacharyya shape (removing the means) by the mediocre performance of the Euclidean and Mahalanobis distances.

The innovations of this research are presented in the following chapter.

CHAPTER VIII

INNOVATIONS

Some of the innovations discovered during the course of this research are shown in Table VIII-1. A synergy was achieved by combining ideas from physiology, speech perception, speech production, statistics, and previous speaker authentication work. As mentioned previously, many speaker verification systems exclude information that contains speaker-dependent information. To meet demanding performance objectives (e.g., the U.S. Government's), speaker discriminatory information cannot be wasted. The speaker discrimination power of perceptually and auditory based features was not successfully demonstrated in this research. Hopefully, future research will uncover effective methods to use these features.

The LSP divergence shape is shown to have strong speaker discriminatory power. The LSP and LP cepstral features were found to be powerful in the divergence measures and Bhattacharyya distances. Numerical limitations precluded the use of sophisticated optimum information-theoretic, linear feature selection techniques. Hopefully, these difficulties can be overcome in future research, as well.

Figure VIII-1 shows some of the basic signal processing blocks used to carry out this research. To process all 20 hours of speech contained in the YOHO database, the computational and disk storage demands of this system are high because of the vast flexibility afforded by this architecture. The storage of all the intermediate processing (e.g., the entire feature set), requires nearly 5 billion bytes (5 GB). The computation of the perceptual model filter bank and the auditory pitch and voicing consumed the equivalent of 3 months of Cray-2 supercomputer CPU time. After this daunting

TABLE VIII-1
INNOVATIONS

-
- **Use excitation information**
 - LP residual
 - Subglottal characteristics
 - Phrase dependent weighting
 - **Additional use of LP information**
 - Singular value decomposition of the LP impulse response matrix
 - Phrase dependent weighting
 - **Perceptually motivated observation set**
 - Perceptually-based filterbank
 - **Speech production features**
 - Line spectrum pair frequencies
 - **Discriminatory measures**
 - Divergence shape
 - Bhattacharyya shape
 - **Statistical methods of feature selection**
 - Divergence
 - Bhattacharyya distance
 - Speaker dependent weighting
 - **Importance sampling**
 - **Risk adaptation**
 - **Data Fusion**
 - **Combine merits of different systems**
-

realization, the author developed a suite of scripts to multiprocess these jobs on a network of 40 Sun SPARC-2 workstation computers. For this processing, the Sun network achieved the throughput of 10 Cray-2 supercomputers! These scripts have revolutionized the way the author's colleagues perform large computational tasks.

The feature and measure found most powerful in this research is the line spectrum pair frequencies feature measured by the divergence shape. Table VIII-2 provides a convenient summary of the performance of a few standard verification systems with the performance of the identification system developed in this research.

TABLE VIII-2
RELATIVE PERFORMANCE

Source	Org	Features	Input	Text	Method	Pop	Error
(Doddington 1985)	TI	Filter-bank	Lab	Dep	DTW	200	~0.8%@6s
(Soong and others 1987)	AT&T	LPC	Phone	Dep (digits)	VQ	100	6%@1s 1.5%@5s
(Higgins and others 1991)	ITT	LAR, LPC Cep	Office	Dep	DTW Likelihood	186	0.7%@20s
Campbell 1992	OSU/DoD	LSP	Office	Indep	Divergence Shape LSPs	44 +43	0.05%@80s

The main contribution is a new information-theoretic shape measure between line spectral pair (LSP) frequency features. This new measure, the divergence shape, can be interpreted geometrically as the shape of an information-theoretic measure called

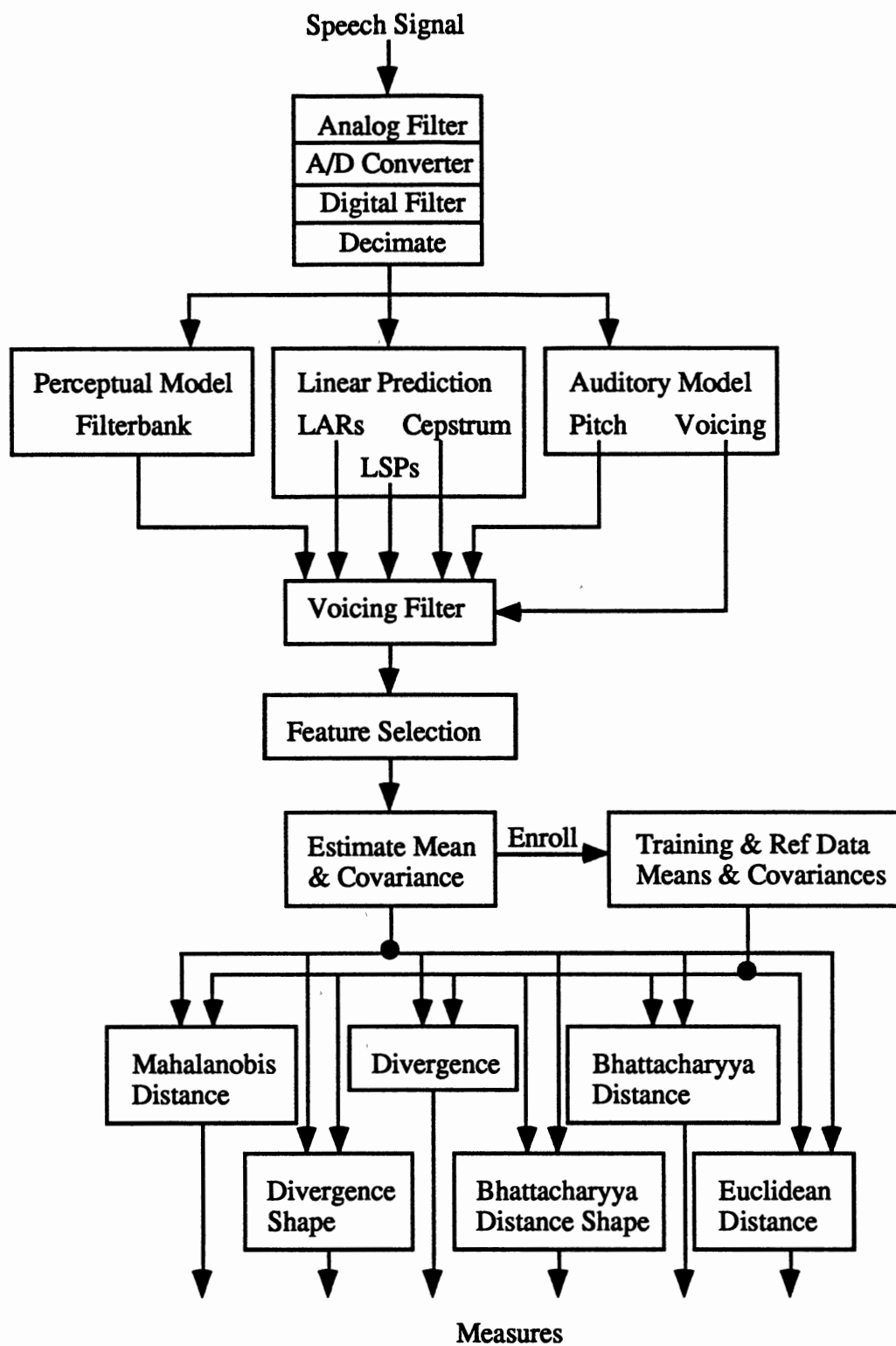


Figure VIII-1. Signal Processing Blocks of New System

divergence. The LSPs were found to be very effective features in this divergence shape measure. This powerful combination will likely become a new standard of reference for future speaker recognition research. The implication of this new measure is vastly improved speaker verification performance relative to the state of the art. This will help save a large portion of the billions of dollars currently lost to telephone credit card fraud annually. Past speaker verification research concentrated in the defense industry and the proposed work will allow for technology transfer to commercial areas.

Accomplishments

The following were developed and implemented by the author in this research LP analysis and conversions (cepstrum, LSPs), perceptual filterbank, and subharmonic summation pitch estimator features; DTW; recursive estimation of mean and covariance; divergence, Bhattacharyya, Mahalanobis and Euclidean measures. An auditory pitch estimator, an HMM, and GSVD were implemented.

A multiprocessing system that revolutionized large computational problems and a linear algebra library optimized for Sun SPARC workstations was developed. Over 6000 lines of documented and verified FORTRAN, C, Bourne shell, MatLab™ and Mathematica™ code were written to conduct the experiments in this research.

Cross-speaker testing (casual impostors) was performed; confusion matrices for each system were generated; wolves and sheep of DTW and NN systems were identified; and weaknesses in features, matching, and models were discovered.

Finally, we are ready to conclude by reviewing the problem at hand, summarizing the major contributions of the research contained in this document, and by suggesting future research.

CHAPTER IX

SUMMARY AND CONCLUSIONS

This work derives and demonstrates new and powerful features and measures for automatic speaker recognition and compares them with traditional ones using speaker discrimination criterion. Automatic speaker recognition is the use of a machine to recognize a person from a spoken phrase. Speaker recognition systems can be used in two modes: to *identify* a particular person or to *verify* a person's claimed identity.

The Problem

The problem under consideration is to discover features and measures to discriminate among individual voices. The scope of this study is limited to speech collected from cooperative users in real-world office environments and without adverse microphone or channel impairments.

Important Findings

In this research, new features and measures for speaker verification were explored and compared with traditional ones using speaker discrimination criterion. It was found that new perceptually based features did not outperform traditional speech production features with respect to speaker identification errors.

Powerful new production features and measures for speaker verification were discovered. The main contribution of this work is a new information-theoretic shape measure between line spectrum pair frequency features. This new measure, the *divergence shape*, can be interpreted geometrically as the shape of an information-

theoretic measure called divergence. The LSPs were found to be very effective features in this divergence shape measure. Experimental results show that these new features and measures yield 0.05% speaker identification error. This is an order of magnitude better than the performance of any other claim reported to date. A speaker verification system using methods presented in this dissertation would be practical to implement in software on a modern personal computer.

Suggestions for Future Research or Study

Additional testing should be performed to increase the statistical confidence level of the experimental results. This would require even greater computational and storage capacity than was used in this work.

For many commercial applications, it's necessary to operate over telephone channels. These channels typically have narrower bandwidths and more noise than the recording conditions of the speech data used in this test (the YOHO database). The features and measures introduced in this research should be evaluated over these channels to test their feasibility for commercial telephone applications.

If a normalization problem can be solved, the generalized singular value decomposition based measures should be further investigated.

The application of these new features and measures in an HMM-based speaker recognizer should be investigated. HMMs offer a powerful way to capture the timing information in speech signals. If this timing information could be combined with the long-term statistical information of the new features and measures (which ignore timing), a very powerful system may result.

The speaker discrimination power of perceptually and auditory based features was not successfully demonstrated in this research. Hopefully, future research will uncover effective methods to use these features.

Numerical limitations precluded the use of sophisticated optimum information-theoretic, linear feature selection techniques. Hopefully, these difficulties can be overcome in future research.

This dissertation has demonstrated the feasibility and power of new features and measures as the front end for a speaker recognition system. To build the back end of a speaker verification system, a method of determining accept/continue/reject thresholds needs to be designed. It's anticipated that the finished system will be able to provide powerful speaker verification because of the strength of those features and measures demonstrated in this dissertation.

CITATIONS

- Atal, B. "A Model of LPC Excitation in terms of the Eigenvectors of the Autocorrelation Matrix of the Impulse Response of the LPC Filter." In *International Conference on Acoustics, Speech, and Signal Processing in Glasgow*, IEEE, 45 – 48, 1989.
- Atal, B. S. "Effectiveness of Linear Prediction Characteristics of the Speech Wave for Automatic Speaker Identification and Verification." *Journal of the Acoustical Society of America* 55, no. 6 (1974): 1304 – 1312.
- Atal, B. S. "Automatic Recognition of Speakers from Their Voices." *Proceedings of the IEEE* 64 (1976): 460 – 475.
- Attili, J., M. Savic, and J. Campbell. "A TMS32020-Based Real Time, Text-Independent, Automatic Speaker Verification System." In *International Conference on Acoustics, Speech, and Signal Processing in New York*, IEEE, 599 – 602, 1988.
- Blahut, R. E. *Principles and Practice of Information Theory*. Electrical and Computer Engineering, Reading: Addison-Wesley, 1987.
- Borden, G. and K. Harris. *Speech Science Primer*. 2nd ed., Baltimore: Williams & Wilkins, 1984.
- Campbell, J. P., Jr. "False Acceptance Errors in Speaker Authentication Systems." Ph.D. Qualifying Examination Report, Oklahoma State University, 1991.
- Campbell, J. P., Jr., T. E. Tremain, and V. C. Welch. "The Federal Standard 1016 4800 bps CELP Voice Coder." *Digital Signal Processing* 1, no. 3 (1991): 145 – 155.
- De Iacovo, R. D., R. Montagna, and D. Sereno. "Vector Quantization and Perceptual Criteria in SVD based CELP Coders." In *International Conference on Acoustics, Speech, and Signal Processing in Albuquerque*, IEEE, 33 – 36, 1990.
- Demmel, J. and W. Kahan. "Accurate Singular Values of Bidiagonal Matrices." *SIAM Journal on Scientific and Statistical Computing* 11, no. 5 (1990): 873 – 912.
- Deprettere, E. F., ed. *SVD and Signal Processing: Algorithms, Applications and Architectures*. Amsterdam: North-Holland, 1988.
- Devijver, P. A. "On a New Class of Bounds on Bayes Risk in Multihypothesis Pattern Recognition." *IEEE Transactions on Computers* C-23, no. 1 (1974): 70 – 80.

- Doddington, G. R. "Speaker Recognition—Identifying People by their Voices." *Proceedings of the IEEE* 73, no. 11 (1985): 1651 – 1664.
- Duda, R. and P. Hart. *Pattern Classification and Scene Analysis*. New York: Wiley, 1973.
- Endsley, J. "Joint Source-Channel Coding with Real Number BCH and Reed-Solomon Codes: Their Properties and Performance in the Presence of Additive Noise." Ph.D. Dissertation, Oklahoma State University, 1991.
- Flanagan, J. *Speech Analysis Synthesis and Perception*. 2nd ed., Berlin: Springer-Verlag, 1972.
- Fukunaga, K. *Introduction to Statistical Pattern Recognition*. 2nd ed., Computer Science and Scientific Computing, ed. W. Rheinboldt and D. Siewiorek. San Diego: Academic Press, 1990.
- Furui, S. "Cepstral Analysis Technique for Automatic Speaker Verification." *IEEE Transactions on Acoustics, Speech, and Signal Processing* ASSP-29, no. 2 (1981): 254 – 272.
- Furui, S. "Speaker-Dependent-Feature Extraction, Recognition and Processing Techniques." *Speech Communication* 10 (1991): 505 – 520.
- Fussell, J. "Speech Processing (52.747) Class Notes." Baltimore: The Johns Hopkins University, 1986.
- Gnanadesikan, R. and J. R. Kettenring. "Discriminant Analysis and Clustering." *Statistical Science* 4, no. 1 (1989): 34 – 69.
- Golub, G. and C. Van Loan. "Matrix Computations." §8.3 and Chapter 12. Baltimore: Johns Hopkins University Press, 1983.
- Harris, F. J. "On the Use of Windows for Harmonic Analysis with the DFT." *Proceedings of the IEEE* 66 (1978): 51 – 83.
- Hedges, L. and I. Olkin. *Statistical Methods for Meta-Analysis*. San Diego: Academic Press, 1985.
- Hermansky, H. "Perceptual Linear Predictive (PLP) Analysis of Speech." *Journal of the Acoustical Society of America* 87, no. 4 (1990): 1738 – 1752.
- Hermes, D. J. "Measurement of Pitch by Subharmonic Summation." *Journal of the Acoustical Society of America* 83, no. 1 (1988): 257 – 264.
- Hermes, D. J. *Pitch Analysis (for Proceedings of ESCA Workshop on Comparing Speech Signal Representations, Sheffield, England, April 7 – 9, 1992)*. Institute for Perception Research, 1992. Manuscript 848.

- Hershey, J. and R. Yarlagadda. "Data Transportation and Protection." 193 – 194. New York: Plenum Press, 1986.
- Higgins, A. "YOHO Speaker Verification." Baltimore: 1990.
- Higgins, A., L. Bahler, and J. Porter. "Speaker Verification Using Randomized Phrase Prompting." *Digital Signal Processing* 1, no. 2 (1991): 89 – 106.
- Higgins, A. L. and R. E. Wohlford. "A New Method of Text-Independent Speaker Recognition." In *International Conference on Acoustics, Speech, and Signal Processing in Tokyo*, IEEE, 869 – 872, 1986.
- Itakura, F. "Line Spectrum Representation of Linear Predictive Coefficients." *Transactions of the Committee on Speech Research, Acoustical Society of Japan* S75 (1975): 34.
- Kailath, T. "The Divergence and Bhattacharyya Distance Measures in Signal Selection." *IEEE Transactions on Communication Technology* COM-15, no. 1 (1967): 52 – 60.
- Kang, G. and L. Fransen. *Low Bit Rate Speech Encoder Based on Line-Spectrum-Frequency*. NRL, 1985. NRL Report 8857.
- Kullback, S. *Information Theory and Statistics*. New York: Dover, 1968.
- Kullback, S. and R. Leibler. "On Information and Sufficiency." *Annals of Mathematical Statistics* 22 (1951): 79 – 86.
- Lee, Y.-T. "Information-Theoretic Distortion Measures for Speech Recognition." *IEEE Transactions on Acoustics, Speech, and Signal Processing* 39, no. 2 (1991): 330 – 335.
- Li, K. P. and E. H. Wrench Jr. "Text-Independent Speaker Recognition with Short Utterances." In *International Conference on Acoustics, Speech, and Signal Processing in Boston*, IEEE, 555 – 558, 1983.
- Makhoul, J. "Linear Prediction: A Tutorial Review." *Proceedings of the IEEE* 63 (1975): 561 – 580.
- Markel, J. D. and S. B. Davis. "Text-Independent Speaker Recognition from a Large Linguistically Unconstrained Time-Spaced Data Base." *IEEE Transactions on Acoustics, Speech, and Signal Processing* ASSP-27, no. 1 (1979): 74 – 82.
- Martin, F. N. *Introduction to Audiology*. 4th ed., Englewood Cliffs: Prentice-Hall, 1991.
- Moler, C., J. Little, and S. Bangert. "MatLab™ for Macintosh Computers." 2-54 – 2-55. South Natick: The MathWorks, Inc., 1989.

- Naik, J. "Speaker Verification: A Tutorial." *IEEE Communications Magazine*, January 1990, 42 – 48.
- Neuburg, E. "A Note on the Frequency Scale." In *Symposium on Acoustic Phonetics and Speech Modeling in Williamstown, Massachusetts*, edited by A. House, IDA/CRD, paper F22, 1981.
- O'Shaughnessy, D. *Speech Communication, Human and Machine*. Digital Signal Processing, Reading: Addison-Wesley, 1987.
- Oppenheim, A. V. and R. W. Schaffer. *Discrete-Time Signal Processing*. Englewood Cliffs: Prentice-Hall, 1989.
- Parsons, T. *Voice and Speech Processing*. Communications and Signal Processing, ed. S. Director. New York: McGraw-Hill, 1987.
- Pentz, A. "Speech Science (SPATH 4313) Class Notes." Stillwater: Oklahoma State University, 1990.
- Press, W., B. Flannery, S. Teukolsky, and W. Vetterling. "Numerical Recipes, The Art of Scientific Computing (FORTRAN version)." 52 – 64. Cambridge: Cambridge University Press, 1990.
- Rabiner, L. and B.-H. Juang. "An Introduction to Hidden Markov Models." *IEEE ASSP Magazine*, January 1986, 4 – 16.
- Rabiner, L. and R. Schaffer. *Digital Processing of Speech Signals*. Signal Processing, ed. A. Oppenheim. Englewood Cliffs: Prentice-Hall, 1978.
- Rabiner, L. R. "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition." *Proceedings of the IEEE* 77, no. 2 (1989): 257 – 286.
- Rosenberg, A. "Automatic Speaker Verification: A Review." *Proceedings of the IEEE* 64, no. 4 (1976): 475 – 487.
- Rosenberg, A. E. and F. K. Soong. "Recent Research in Automatic Speaker Recognition." In *Advances in Speech Signal Processing*, ed. S. Furui and M. M. Sondhi. 701 – 738. New York: Marcel Dekker, 1992.
- Saito, S. and K. Nakata. *Fundamentals of Speech Signal Processing*. Tokyo: Academic Press, 1985.
- Sakoe, H. and S. Chiba. "Dynamic Programming Algorithm Optimization for Spoken Word Recognition." *IEEE Transactions on Acoustics, Speech, and Signal Processing* ASSP-26, no. 1 (1978): 43 – 49.
- Sanchez-Calle, V., J. Lopez-Soler, J. Segura-Luna, A. Peinado-Herreros, and A. Rubio-Ayuso. "Increasing the Difference between the Significant and the Non-Significant Singular Values in a Model of LPC Excitation Based on the SVD." In

Fifth European Signal Processing Conference in Barcelona, Elsevier, 1287 – 1290, 1990.

- Schwartz, R., S. Roucos, and M. Berouti. "The Application of Probability Density Estimation to Text Independent Speaker Identification." In *International Conference on Acoustics, Speech, and Signal Processing in Paris*, IEEE, 1649 – 1652, 1982.
- Slaney, M. *Lyon's Cochlear Model*. Apple Computer, Inc., 1988. Apple Technical Report 13.
- Soong, F. K., A. E. Rosenberg, L. R. Rabiner, and B.-H. Juang. "A Vector Quantization Approach to Speaker Recognition." *AT&T Technical Journal* 66, no. 2 (1987): 14 – 26.
- Strang, G. *Linear Algebra and its Applications*. 3rd ed., San Diego: Harcourt Brace Jovanovich, 1988.
- Sutherland, A. and M. Jack. "Speaker Verification." In *Aspects of Speech Technology*, ed. M. Jack and J. Laver. 185 – 215. Edinburgh: Edinburgh University Press, 1988.
- Tishby, N. Z. "On the Application of Mixture AR Hidden Markov Models to Text Independent Speaker Recognition." *IEEE Transactions on Acoustics, Speech, and Signal Processing* 39, no. 3 (1991): 563 – 570.
- Tobias, J. V. *Foundations of Modern Auditory Theory*. New York: Academic Press, 1970.
- Tou, J. and R. Gonzalez. *Pattern Recognition Principles*. Applied Mathematics and Computation, ed. R. Kalaba. Reading: Addison-Wesley, 1974.
- Tou, J. and P. Heydorn. "Some Approaches to Optimum Feature Extraction." In *Computer and Information Sciences-II*, ed. J. Tou. 57 – 89. COINS II. New York: Academic Press, 1967.
- Trancoso, I. and B. Atal. "Efficient Search Procedures for Selecting the Optimum Innovation in Stochastic Coders." *IEEE Transactions on Acoustics, Speech, and Signal Processing* 38, no. 3 (1990): 385 – 396.
- Vaccaro, R. J., ed. *SVD and Signal Processing, II: Algorithms, Analysis and Applications*. Amsterdam: Elsevier, 1991.
- Van Immerseel, L. M. and J.-P. Martens. "Pitch and Voiced/Unvoiced Determination with an Auditory Model." *Journal of the Acoustical Society of America* 91, no. 6 (1992): 3511 – 3526.
- Vetterling, W., S. Teukolsky, W. Press, and B. Flannery. "Numerical Recipes, Example Book (FORTRAN version)." Chapters 2 and 14. Cambridge: Cambridge University Press, 1989.

Wald, A. *Sequential Analysis*. New York: Wiley, 1947.

Whalen, A. *Detection of Signals in Noise*. Electrical Science, ed. H. Booker and N. DeClaris. New York: Academic Press, 1971.

Wolfram, S. "Mathematica™, A System for Doing Mathematics by Computer." 454 – 455. Redwood City: Addison-Wesley, 1988.

Yarlagadda, R. "Personal Communication." 1991.

BIBLIOGRAPHY

Although the following works are not explicitly cited in this dissertation, they were influential in this research.

- Anderson, T. W. *An Introduction to Multivariate Statistical Analysis*. 2nd ed., A Wiley Publication in Mathematical Statistics, New York: Wiley, 1984.
- Atal, B. S. "Automatic Speaker Recognition Based on Pitch Contours." *Journal of the Acoustical Society of America* 52, no. 6 (1972): 1687 – 1697.
- Atal, B. S., J. L. Miller, and R. D. Kent, ed. *Papers in Speech Communication: Speech Processing*. Vol. 3. Papers in Speech Communication. Woodbury: Acoustical Society of America, 1991.
- Attili, J. B. "On the Development of a Real-Time Text-Independent Speaker Verification System." Ph.D. Dissertation, Rensselaer Polytechnic Institute, 1987.
- Barbosa, L. C. "A Maximum-Energy-Concentration Spectral Window." *IBM Journal of Research and Development* 30, no. 3 (1986): 321 – 325.
- Barnes, E. R. "An Algorithm for Separating Patterns by Ellipsoids." *IBM Journal of Research and Development* 26, no. 6 (1982): 759 – 764.
- Bogner, R. E. "On Talker Verification Via Orthogonal Parameters." *IEEE Transactions on Acoustics, Speech, and Signal Processing* ASSP-29, no. 1 (1981): 1 – 12.
- Bolt, R. H., F. S. Cooper, E. E. David Jr., P. B. Denes, J. M. Pickett, and K. N. Stevens. "Speaker Identification by Speech Spectrograms: A Scientist's View of its Reliability for Legal Purposes." *Journal of the Acoustical Society of America* 47, no. 2 (1970): 597 – 612.
- Bolt, R. H., F. S. Cooper, D. M. Green, S. L. Hamlet, J. G. McKnight, J. M. Pickett, O. I. Tosi, B. D. Underwood, D. L. Hogan, and W. Banks. *On the Theory and Practice of Voice Identification*. Washington: National Academy of Sciences, 1979.
- Chen, C. H. *Statistical Pattern Recognition*. Rochelle Park: Hayden, 1973.
- Chen, C. H., ed. *Digital Waveform Processing and Recognition*. Boca Raton: CRC Press, 1982.

- Cover, T. M. and J. A. Thomas. *Elements of Information Theory*. Telecommunications, ed. Donald L. Schilling. New York: Wiley, 1991.
- Crochiere, R. E. and L. R. Rabiner. *Multirate Digital Signal Processing*. Prentice-Hall, 1983.
- Crystal, T. H. *Speaker Authentication Monitoring: Doomed to Failure?* IDA/CRD, 1990.
- Denes, P. and E. Pinson. *The Speech Chain*. New York: Doubleday, 1973.
- Dixon, N. R. and T. B. Martin, ed. *Automatic Speech & Speaker Recognition*. New York: IEEE Press, 1979.
- Fallside, F. and W. Woods, ed. *Computer Speech Processing*. London: Prentice/Hall International, 1985.
- Fant, G. *Acoustic Theory of Speech Production: With Calculations based on X-Ray Studies of Russian Articulations*. Description and Analysis of Contemporary Standard Russian, The Hague: Mouton, 1970.
- Foley, D. and J. Sammon. "An Optimal Set of Discriminant Vectors." *IEEE Transactions on Computers* c-24, no. 3 (1975): 281 – 289.
- Gish, H., K. Karnofsky, M. Krasner, S. Roucos, W. Russell, R. Schwartz, and J. Wolf. *ISIS Literature Survey*. BBN, 1986. Task Report 6142.
- Haykin, S. *Adaptive Filter Theory*. 2nd ed., Englewood Cliffs: Prentice-Hall, 1991.
- Hess, W. *Pitch Determination of Speech Signals: Algorithms and Devices*. Information Sciences, ed. M. Schroeder. Berlin: Springer-Verlag, 1983.
- Higgins, A. "Tutorial Papers." 1988.
- Higgins, A. and J. Porter. *YOHO Speaker Authentication Final Report*. ITT Defense Communications Division, 1989.
- Holmes, M. H. and L. A. Rubenfeld, ed. *Mathematical Modeling of the Hearing Process*. Lecture Notes in Biomathematics. Berlin: Springer-Verlag, 1980.
- House, A. S. *The Recognition of Speech by Machine—A Bibliography*. IDA/CRD, 1989. CRD Technical Report 28.
- Jesorsky, P., ed. *Principles of Automatic Speaker-Recognition*. Speech Communication with Computers. New York: Macmillan, 1978.
- Juang, B.-H. "Hidden Markov Model and its Application to Speech Processing (MA-513) Class Notes." Linthicum: AT&T Bell Labs, 1988.

- Krasner, M., H. Gish, J. Makhoul, S. Roucos, and R. Schwartz. *YOHO Speaker Authentication, Part II: Technical Proposal*. BBN Laboratories, 1986. Proposal P86-CISD-010.
- Kullback, S., J. C. Keegel, and J. H. Kullback. *Topics in Statistical Information Theory*. Vol. 42. Lecture Notes in Statistics, ed. D. Brillinger, S. Fienberg, J. Gani, J. Hartigan, and K. Krickeberg. New York: Springer-Verlag, 1987.
- Ladefoged, P. *Elements of Acoustic Phonetics*. Chicago: The University of Chicago Press, 1962.
- Lass, N., L. McReynolds, J. Northern, and D. Yoder, ed. *Speech, Language, and Hearing*. Vol. 1, Normal Processes. Philadelphia: W. B. Sanders, 1982.
- Lewis, F. L. *Optimal Estimation: With an Introduction to Stochastic Control Theory*. A Wiley-Interscience publication, New York: Wiley, 1986.
- Markel, J. and A. Gray. *Linear Prediction of Speech*. Communication and Cybernetics, Berlin: Springer-Verlag, 1976.
- Markel, J. D., B. T. Oshika, and A. H. Gray Jr. "Long-Term Feature Averaging for Speaker Recognition." *IEEE Transactions on Acoustics, Speech, and Signal Processing ASSP-25* (1977): 330 – 337.
- Myers, D., W. D. Schlosser, R. J. Wolfson, R. A. Winchester, and N. Carmel. *Otologic Diagnosis and the Treatment of Deafness*. Summit: CIBA Pharmaceutical Company, 1970.
- Offer, E., D. Malah, and A. Dembo. "A Unified Framework for LPC Excitation Representation in Residual Speech Coders." In *International Conference on Acoustics, Speech, and Signal Processing in Glasgow*, IEEE, 41 – 44, 1989.
- Oppenheim, A. and R. Schaffer. *Digital Signal Processing*. Englewood Cliffs: Prentice-Hall, 1975.
- Paige, C. C. "Computing the Generalized Singular Value Decomposition." *SIAM Journal on Scientific and Statistical Computing* 7, no. 4 (1986): 1126 – 1146.
- Papoulis, A. *The Fourier Integral and its Applications*. New York: McGraw-Hill, 1962.
- Papoulis, A. *Probability, Random Variables, and Stochastic Processes*. 2nd ed., New York: McGraw-Hill, 1984.
- Press, W. *Wavelet Transforms (to appear in Numerical Recipes: The Art of Scientific Computing, 2nd ed.)*. Harvard-Smithsonian Center for Astrophysics, 1991. Preprint 3184.

- Rosenberg, A. E. and F. K. Soong. "Evaluation of a Vector Quantization Talker Recognition System in Text Independent and Text Dependent Modes." *Computer Speech and Language* 22 (1987): 143 – 157.
- Sanders, D. A. *Auditory Perception of Speech*. Englewood Cliffs: Prentice-Hall, 1977.
- Saunders, W. H. *The Larynx*. Summit: CIBA Pharmaceutical Company, 1964.
- Scharf, L. *Statistical Signal Processing: Detection, Estimation, and Time Series Analysis*. Digital Signal Processing, ed. R. Roberts. Reading: Addison-Wesley, 1991.
- Schroeder, M., B. Atal, and J. Hall. "Objective Measure of Certain Speech Signal Degradations Based on Masking Properties of Human Auditory Perception." In *Frontiers of Speech Communication Research*, ed. Lindblom and Ohman. London: Academic Press, 1979.
- Schwab, E. C. and H. C. Nusbaum, ed. *Pattern Recognition by Humans and Machines*. Vol. 1. Cognition and Perception. San Diego: Academic Press, 1986.
- Sekey, A. and B. Hanson. "Improved One-Bark Bandwidth Auditory Filter." *Journal of the Acoustical Society of America* 75, no. 6 (1984): 1902 – 1904.
- Slaney, M. and R. Lyon. "A Perceptual Pitch Detector." In *International Conference on Acoustics, Speech, and Signal Processing in Albuquerque*, IEEE, 357 – 360, 1990.
- Soong, F. K. and A. E. Rosenberg. "On the Use of Instantaneous and Transitional Spectral Information in Speaker Recognition." *IEEE Transactions on Acoustics, Speech, and Signal Processing* 36, no. 6 (1988): 871 – 879.
- Spicer, C. C. "Calculation of Power Sums of Deviations about the Mean." *Applied Statistics* 21, no. 2 (1972): 226 – 227.
- Stevens, S. S. and H. Davis. *Hearing: Its Psychology and Physiology*. New York: Wiley, 1966.
- Timcke, R. H., H. von Leden, and P. Moore. "Laryngeal Vibrations: Measurements of the Glottic Wave." *Archives of Otolaryngology* 68 (1958): 1 – 19.
- Turabian, K. L. *A Manual for Writers of Term Papers, Theses, and Dissertations*. 5th ed., Chicago: The University of Chicago Press, 1987.
- Vaidyanathan, P. P. "Multirate Digital Filters, Filter Banks, Polyphase Networks, and Applications: A Tutorial." *Proceedings of the IEEE* 78, no. 1 (1990): 56 – 93.
- Van Loan, C. "Computing the CS and Generalized Singular Value Decomposition." *Numerische Mathematik* 46 (1985): 479 – 492.
- von Békésy, G. *Experiments in Hearing*. Translated by E. G. Wever. ed. E. G. Wever. New York: McGraw-Hill, 1960.

- Wang, S. "Low Bit-Rate Vector Excitation Coding of Phonetically Classified Speech." Ph.D. Dissertation, University of California at Santa Barbara, 1991.
- Wang, S., A. Sekey, and A. Gersho. "Auditory Distortion Measure for Speech Coding." In *International Conference on Acoustics, Speech, and Signal Processing in Toronto*, IEEE, 493 – 496, 1991.
- Widrow, B. and S. D. Stearns. *Adaptive Signal Processing*. Englewood Cliffs: Prentice-Hall, 1985.
- Wohlford, R. E., E. H. Wrench Jr., and B. P. Landell. "A Comparison of Four Techniques for Automatic Speaker Recognition." In *International Conference on Acoustics, Speech, and Signal Processing in Denver*, IEEE, 908 – 911, 1980.
- Yarlagadda, R. "Data Transportation and Protection (ECEN 5543) Class Notes." Stillwater: Oklahoma State University, 1991.

2
VITA

Joseph Paul Campbell, Jr.

Candidate for the Degree of

Doctor of Philosophy

Thesis: FEATURES AND MEASURES FOR SPEAKER RECOGNITION

Major Field: Electrical Engineering

Biographical:

Personal Data: Born in Oneonta, New York, December 20, 1956, the son of Joseph Campbell, Sr. and Sally Fionte.

Education: Graduated from Oneonta Senior High School, Oneonta, New York, in 1975; received Associate in Science Degree in Engineering Science from Broome Community College in 1977; received Bachelor of Science Degree in Electrical Engineering from Rensselaer Polytechnic Institute in 1979; received Master of Science Degree in Electrical Engineering from The Johns Hopkins University in 1986; completed requirements for the Doctor of Philosophy degree at Oklahoma State University in 1992.

Professional Experience: Mr. Campbell was with the U.S. Department of Defense's Narrow-Band Secure Voice Technology research group from 1979 to 1990. Since 1984, Mr. Campbell has directed university and industry contracts in speech coding, voice verification, and noise-canceling microphone design. In 1986, he delivered LPC-10e, which redefined the Government Standard and the state of the art in 2400 bps speech coding. In 1988, Mr. Campbell led the U.S. Government's speech coding team in the development of the 4800 bps CELP voice coder that won a Consortium test and became Federal Standard 1016. Since 1991, Mr. Campbell has been with the U.S. Department of Defense's Biometric Authentication Technology research group. His current research activities include speech coding and voice verification. Mr. Campbell has chaired sessions at the IEEE International Conferences on Acoustics, Speech, and Signal Processing. He is a charter associate editor for the newly formed IEEE Transactions on Speech and Audio Processing. Mr. Campbell is a member of The Acoustical Society of America; The Audio Engineering Society; The IEEE; The IEEE Communications Society; The IEEE Acoustics, Speech and Signal Processing Society; The Speech Technical Committee of the IEEE ASSP Society; and Sigma Xi.