

**Analysis of Variability in Speech
with Applications
to Speech and Speaker Recognition**

Sachin S. Kajarekar

B.E., Bombay University, Bombay, 1994

M.S., Oregon Graduate Institute of Science and Engineering, 1998

A dissertation submitted to the faculty of the
OGI School of Science and Engineering
Oregon Health and Science University
in partial fulfillment of the
requirements for the degree
Doctor of Philosophy
in
Computer Science and Engineering

July 2002

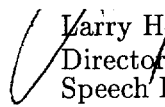
© Copyright 2002 by Sachin S. Kajarekar
All Rights Reserved

The dissertation “Analysis of Variability in Speech with Applications to Speech and Speaker Recognition” by Sachin S. Kajarekar has been examined and approved by the following Examination Committee:


Hynek Hermansky
Professor
Thesis Research Adviser

Xubo Song
Assistant Professor

Bishnu Atal


Larry Heck
Director
Speech Research and Development
Nuance Communications

Misha Pavel
Professor

Dedication

To Aaji, Aajoba, Aai, Baba, and Sona

Acknowledgments

My graduate and doctoral studies were the first steps of my academic career. They were successful because of help and motivation from a lot of people. First of all, I want to thank Late Prof. Michael Macon. He gave me research assistantship during the first quarter of my masters studies. He also gave a good recommendation to my thesis adviser. His feedback during the initial part of this work was invaluable.

I want to thank my thesis adviser, Prof. Hynek Hermansky, for the discussions and feedback on this work. I also want to thank him providing financial support throughout this work. I learned a lot from him about presentation skills, writing skills, research methodology, and so on. I will always remember his anecdotes about the history of speech recognition, his comments of abbreviations (CMS = cumulative multiple sclerosis) and his requirements of a good presentation (why, what and how).

I want to thank my thesis committee members – Prof. Xubo Song, Bishnu Atal, Larry Heck and Prof. Misha Pavel. Their feedback on the thesis has significantly improved the quality of this work. I also want to thank Prof. Jan Van Santen for his comments on this work, specifically on ANOVA.

During the masters and doctoral work, I was fortunate to learn about various topics from different professors – Hynek Hermansky, Michael Macon, Eric Wan, Robert Jaffe, Peter Heeman, Todd Leen, Andrew Fraser, B. Yegnanarayana, David Maier and so on. I want to thank them for broadening my perspective of speech research and implanting critical thinking.

I was fortunate to work with wonderful people in my group and also in Center for Spoken Language Understanding. I want to thank Naren for the numerous discussion, from cricket to speech research, during the early stages of my thesis. In fact, the idea of analyzing variability in speech was a result of one on those long discussions. I want

to thank Ravi and Sangita for their constant support and encouragement. Pratibha and Sunil deserve special thanks not only for technical discussions but also for bearing with me and my cooking experiments during our stay at Apt 26. I also thank Pavel and Alena for their help in computing and administration facilities. I want to thank Andre for reviewing the thesis draft(s). Ed Kaiser needs a special mention for his help in shorting the thesis "title". I want to thank Cynthia Archer for her help in the derivations.

I want to thank my parents for supporting my decision to pursue doctoral studies. They have been a constant source of help and motivation throughout my life (and this work). In addition to making small discoveries in the field of speech and speaker recognition, I was also fortunate to discover my life partner, Sonali. She has been with me on numerous emotional roller-coaster rides, to encourage me, to comfort me, and to cook great meals for me. I want to thank her for all her help and want to assure her that this is the last weekend when I will be working on my thesis.

Contents

| | |
|--------------------------------------------------------|-----------|
| Dedication | iv |
| Acknowledgments | v |
| Abstract | xv |
| 1 Introduction | 1 |
| 1.1 Speech and Speaker Recognition Systems | 2 |
| 1.2 Feature Extraction | 4 |
| 1.2.1 Feature Characteristics | 4 |
| 1.2.2 Conventional Feature Extraction | 5 |
| 1.3 Variability in Features | 9 |
| 1.3.1 Phone (or inter-phone) Variability | 11 |
| 1.3.2 Speaker Variability | 11 |
| 1.3.3 Channel Variability | 11 |
| 1.3.4 Residual Variability | 12 |
| 1.4 Contributions of the Work | 12 |
| 1.5 Organization of Thesis | 14 |
| 2 Analysis of Variability in Speech | 16 |
| 2.1 ANOVA | 16 |
| 2.2 MANOVA | 18 |
| 2.3 Assumptions of MANOVA | 21 |
| 2.4 Database and Features | 23 |
| 2.5 Nature of Variability in Spectral Domain | 25 |
| 2.6 Nature of Variability in Temporal Domain | 27 |
| 2.7 Contribution of Sources of Variability | 29 |
| 2.8 Results in Perspective | 31 |
| 2.8.1 Hierarchical ANOVA | 33 |
| 2.8.2 Mutual Information | 34 |
| 2.9 Conclusions | 40 |

| | | |
|----------|---------------------------------------------------------|-----------|
| 3 | MANOVA as a Diagnostic Tool | 42 |
| 3.1 | Speaker Variability and Speaker Recognition Performance | 42 |
| 3.1.1 | Feature-set | 43 |
| 3.1.2 | MANOVA | 44 |
| 3.1.3 | Speaker Recognition Experiments | 45 |
| 3.2 | Speaker Recognition using Broad Phonetic Categories | 47 |
| 3.2.1 | Speaker Recognition Task | 47 |
| 3.2.2 | Recognition System | 48 |
| 3.2.3 | Results | 48 |
| 3.3 | Phone Variability and Speech Recognition Performance | 50 |
| 3.3.1 | Effect of Delta and Double-delta Features | 51 |
| 3.3.2 | Effect of Feature Transformations | 52 |
| 3.4 | Conclusions | 53 |
| 4 | MANOVA and LDA | 55 |
| 4.1 | MANOVA | 55 |
| 4.2 | Linear Discriminant Analysis | 56 |
| 4.3 | LDA in Speech Recognition | 57 |
| 4.3.1 | Spectral Domain | 57 |
| 4.3.2 | Temporal Domain | 58 |
| 4.4 | Effect of Phone Classes on LDA | 58 |
| 4.5 | Three-state MANOVA | 60 |
| 4.6 | Three-state LDA | 62 |
| 4.7 | Conclusions | 65 |
| 5 | Two-Dimensional Discriminants for ASR | 67 |
| 5.1 | Two-dimensional Linear Discriminants | 68 |
| 5.1.1 | Joint Analysis | 69 |
| 5.1.2 | Combined Analysis | 69 |
| 5.2 | Optimality of Combined Discriminants | 71 |
| 5.3 | 2D Discriminants in Speech | 75 |
| 5.4 | Data Insufficiency for Joint Discriminants | 76 |
| 5.5 | Conclusions | 79 |
| 6 | Summary and Conclusions | 80 |
| 6.1 | Summary | 80 |
| 6.2 | Future Directions | 83 |
| 6.3 | Final Comments | 84 |

| | | |
|----------|---------------------------------------------------------|------------|
| A | Phone Set for MANOVA | 86 |
| B | Effect of Phonetic Context on Results of MANOVA | 87 |
| C | Temporal Variability in Different Bands | 89 |
| D | Temporal Filter for Speaker Verification | 91 |
| E | Different Measures for Source Contributions | 94 |
| F | Line Spectral Pair and MFC Coefficients | 99 |
| G | Analysis of Variability using Mutual Information | 102 |
| | G.1 Experimental Setup | 103 |
| | G.2 MANOVA | 103 |
| | G.3 Information-theoretic Analysis | 105 |
| | G.4 Results in Perspective | 109 |
| | G.5 Conclusions | 110 |
| | Bibliography | 111 |
| | Biographical Note | 116 |

List of Tables

| | | |
|-----|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----|
| 2.1 | Contribution of sources in spectral domain (a) and temporal domain (b) in HTIMIT database. Numbers in the bracket indicate the variation in the results over 7 different data-sets. | 32 |
| 2.2 | % contribution of sources in spectral domain without (a) and with UMS (b) | 32 |
| 2.3 | % contribution of sources in temporal domain without (a) and with UMS (b) | 33 |
| 3.1 | Phones and their mapping to broad phonetic categories | 44 |
| 3.2 | Relative speaker variability for features of different broad categories | 45 |
| 3.3 | % equal error rates (EERs) using different broad phonetic categories. Same handset condition refers to trials where same handset was used for training and testing. Different handset condition refers to trials where different handset was used for training and testing. | 47 |
| 3.4 | Speaker recognition performance using broad phonetic categories on NIST 2001 speaker recognition task | 49 |
| 3.5 | Speaker recognition performance using combination of broad phonetic categories on NIST 2001 speaker recognition task | 49 |
| 3.6 | Phone variability and speech recognition performance using delta and double-delta features | 51 |
| 3.7 | Phone variability and recognition performance at different stages of proposed AURORA front-end | 54 |
| 4.1 | Contribution of variabilities in spectral domain using 1-state and 3-state phone models | 62 |
| 4.2 | Contribution of variabilities in temporal domain using 1-state and 3-state phone models | 62 |
| 4.3 | % WER using one-state and three-state spectral discriminants (SLDs) and temporal discriminants (TLDs) | 65 |
| 5.1 | Connected digit recognition performance using 24 discriminant features. . . . | 76 |
| 5.2 | Effect of different types of TLDs on CLDs | 77 |
| A.1 | Phone set common between OGI Stories, HTIMIT and TIMIT databases. . . . | 86 |

| | | |
|-----|----------------------------------------------------------------------------------------------------------------------------|-----|
| B.1 | Contribution of variabilities using context-independent (CI) phones and context-dependent (CD) phones as classes | 88 |
| E.1 | Different measurements using MANOVA and corresponding recognition performance | 96 |
| F.1 | Comparison of source contributions using MANOVA using LSP and MFCC features. | 100 |
| G.1 | Contribution of sources in spectral and temporal domains | 105 |
| G.2 | Mutual information between features and phone and speaker and channel labels in spectral and temporal domains | 107 |

List of Figures

| | | |
|-----|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----|
| 1.1 | Block diagram of speech and speaker recognition systems | 3 |
| 1.2 | Speech samples as features | 9 |
| 1.3 | Power spectral energies as features | 9 |
| 1.4 | Bark-warped power spectral energies as features | 10 |
| 1.5 | Bark-warped power spectral energies with logarithmic compression as features. | 10 |
| 1.6 | Discrete cosine transform coefficients as features | 10 |
| 2.1 | Sources of variability in speech using MANOVA | 19 |
| 2.2 | Spectral and temporal feature vectors used for MANOVA | 25 |
| 2.3 | Difference in the computation of phone variability in spectral and temporal domains | 26 |
| 2.4 | MANOVA in spectral domain using HTIMIT database. Before utterance-based mean subtraction (MS) (a) and after UMS (b). Phone information (solid line) (I), speaker information (dotted line), channel information (dash-dot line), and error information (dash line). Note the reduction in speaker and channel information after UMS. | 27 |
| 2.5 | MANOVA in temporal domain using HTIMIT database. Before utterance-based mean subtraction (MS) (a) and after UMS (b). Phone variability (solid line), speaker variability (dotted line), channel variability (dash-dot line), and error variability (dash line). Note the reduction in the speaker and channel variability after UMS. | 28 |
| 2.6 | Mutual information between frequency band and the phone labels computed assuming Gaussian distribution | 37 |
| 2.7 | Mutual information between feature at time t and the phone label at 5 Barks assuming Gaussian distribution | 38 |
| 2.8 | Joint mutual information between two features in frequency and the phone labels given that first measurement in frequency is at 4 or 5 Barks assuming Gaussian distribution | 39 |

| | | |
|-----|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----|
| 2.9 | Joint mutual information between two measurements in time and the phone labels given that first measurement in time is at $t = 0$. Results are frequency of 5 Barks assuming Gaussian distribution. | 39 |
| 3.1 | Speaker and channel variability within each phone. | 46 |
| 3.2 | Universal background model - Gaussian mixture model (UBM-GMM) framework | 47 |
| 3.3 | Performance of speaker recognition system using all categories (dotted-line) and two broad phone categories (straight-line) on NIST 2001 speaker verification task | 50 |
| 3.4 | Modified QUALCOMM-OGI-ICSI front-end used for AURORA task | 52 |
| 4.1 | Effect of separation of closure and burst on the design of TLDs. | 59 |
| 4.2 | Block diagram of HMM where states follow each other in a sequence. . . . | 60 |
| 4.3 | Phone variance (thin line) and average within-state variance (thick line) in spectral and temporal domain | 63 |
| 4.4 | Effect of 3-state design on TLDs | 64 |
| 5.1 | Procedure for obtaining JLDs. | 69 |
| 5.2 | Procedure for obtaining CLDs. | 71 |
| 5.3 | The First 8 LDs from spectral domain | 72 |
| 5.4 | The 3 LDs from temporal trajectory of first spectral discriminant feature . | 72 |
| 5.5 | CLDs obtained by combining first two SLDs and TLDs. The title indicates the index of the SLD and TLD. | 73 |
| 5.6 | Insufficient training data for JLDs. (a) shows that JLDs need more training data than CLDs. (b) shows that in absence of sufficient training data JLDs over-fit the training data. | 78 |
| B.1 | Results of MANOVA using context-independent phones and context-dependent phones as classes in spectral domain | 88 |
| B.2 | Results of MANOVA using context-independent and context-dependent phones as classes in temporal domain | 88 |
| C.1 | Temporal variability in bands 1-3: phone variability (thick solid line), speaker variability (thin solid line), channel variability (thick dot-dash line), residual variability (thick dotted line). Number in each plot is the critical band index. | 89 |

| | | |
|-----|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----|
| C.2 | Temporal variability in bands 4-15:phone variability (thick solid line), speaker variability (thin solid line), channel variability (thick dot-dash line), residual variability (thick dotted line). Number in each plot is the critical band index. | 90 |
| D.1 | Impulse and frequency response of the filter used in speaker recognition system | 92 |
| D.2 | Performance of old and new filter on speaker verification | 93 |
| E.1 | Difference between %PHN1 and %PHN2. For Case (I) and Case (II), %PHN1 will be same because the marginal distributions are same. %PHN2 for these cases will be different because, in the joint space, classes have less overlap in Case (II) than Case (I). | 97 |
| G.1 | Results of analysis of variability | 105 |
| G.2 | Results of information-theoretic analysis | 107 |

Abstract

Analysis of Variability in Speech with Applications to Speech and Speaker Recognition

Sachin S. Kajarekar

Supervising Professor: Hynek Hermansky

The speech signal has variability due to language, speakers, and communication channels. In this work, variability due to language is referred to as variability due to different phones in the language. It is also referred to as (inter-phone or) phone variability. Variability due to speakers is referred to as speaker variability, and variability due to different communication channels is referred as channel variability. The remaining variability in the signal is referred to as residual variability.

The total variability in speech is decomposed using multivariate analysis of variance (MANOVA). Here variability in speech refers to variability in the set of features extracted from speech signal, and variability refers to covariance of features due to different phones, different speakers, and different channels. In this work, MANOVA is performed using three databases - HTIMIT, OGI Stories and OGI Numbers. Variability in the commonly used features is measured in spectral and temporal domains. The results are shown to be consistent across different databases and datasets. The results are also shown to be consistent with the previous studies.

The results of MANOVA are applied in two ways. First, we show that contribution of the variabilities in features is related to their performance on speech and speaker

recognition tasks. Second, we show that results of MANOVA can be used for deriving discriminant features for a given task.

Relationship between results of MANOVA and speech and speaker recognition results is illustrated using several examples. First, the speaker variability in different broad phonetic categories is computed using MANOVA. It is shown that the results of the analysis are related to the performance of speaker recognition system using these categories. Second, we compare contribution of phone variability in different types of feature-sets, for example, features with delta and double delta features, features after RelAtive SpecTRAl (RASTA) filtering, features after concatenation with TempoRAI PatternS (TRAPS) features and so on. We show that the change in the contribution of phone variability is related to change in the performance of the features on speech recognition task.

Using MANOVA, we had observed that the variability due to phones spreads for approximately 250 ms around the current frame. We include this variability in the design of features using Linear discriminant analysis (LDA). Two types of analysis are performed. First analysis, called joint analysis, uses all the correlations in a block of spectrogram. Second analysis, called combined analysis, assumes that time and frequency domains are independent. The discriminant features from both analysis are used in speech recognition experiments. The results show that features from joint analysis perform worse than combined analysis because joint analysis over-fits the training data and does not generalize on the test data. In general, we show that performance of the speech recognition system improves when information from a longer time-span is included in the features. This is constant with the result of MANOVA. Specifically, we show that combination of spectral and temporal discriminants yields to the best joint time-frequency discriminants.

Chapter 1

Introduction

Speech communication is a primary form of human-to-human communication. In this form, the speech signal is used to communicate linguistic messages. However, the signal also carries information about the characteristics of speaker and communication channel. Due to naturalness of the speech interface, researchers have investigated its use for human-computer interaction. This thesis deals with a part of the interface that recognizes the speech and the speaker.

Speech recognition is a process of converting the speech signal to sequence of words. The words are chosen from a dictionary. The sequence of words is determined using a language model. Speech recognition systems are characterized by many parameters [45] such as size of vocabulary, language model, perplexity, speaking style, speaking mode. Commonly used speech recognition systems are dictation systems and digit recognition systems. Dictation systems have a large vocabulary (≥ 20000 words) and higher perplexity (≥ 100). Digit recognition systems have small vocabulary (≈ 11 words) and lower perplexity (≤ 5).

Speaker recognition is a process of recognizing the speaker from his or her voice [11, 21]. The process can be performed in two modes – close-set or open-set. In close-set mode, the test speaker is one of the speakers in the training set. In open-set mode, the test speaker may or may not be the one from the training set. There are two application of speaker recognition – speaker identification and speaker verification. Speaker identification refers to speaker recognition within a set of speakers, which is also referred to as a closed-set task. In this case, the utterance is tested against all the speakers in the set, and the speaker with the highest likelihood is declared as the hypothesized speaker. The applications of

speaker identification are in forensic studies where a speech evidence is used to recognize the identity of a known criminal. Speaker verification needs the claimed identity of the speaker along-with the speech sample. The task is to verify the claimed identity of the speaker based on his voice. This is an open-set task because the test speaker may or may not belong to the closed-set. The applications of speaker verification are in the security area where the speaker characteristics are used as voice-signature.

The most common cause of performance degradation of the speech and speaker recognition systems is a mismatch between training and testing conditions. For example, when a speech recognition system is trained under noise-free condition, and it is used inside a running car. In general, the mismatch can be compensated at two levels: feature extraction and model estimation. In this work, we address the mismatch at the feature extraction level. We study the nature and the contribution of different types of variability in speech, and use the results of the study to improve the performance of the systems under mismatch conditions.

The chapter is organized as follows. Section 1.1 gives a brief overview of automatic speech recognition (ASR) and speaker recognition systems. A common processing step in these systems is the estimation of a set of parameters from the speech signal. This is referred to as features extraction. This is described in detail in Section 1.2. The variability in the features can be attributed to different type of variabilities in speech. Section 1.3 describes four types of variabilities used in this work, and their usefulness for the speech and speaker recognition task. The chapter concludes with contributions of the thesis in Section 1.4, and overview of the thesis in Section 1.5.

1.1 Speech and Speaker Recognition Systems

Speech and speaker recognition systems are based on a pattern recognition framework [12, 18]. Figure 1.1 shows the block diagram of these systems. They have three main processing steps - feature extraction, likelihood computation, and search for the most likely output (decoder). Feature extraction step is implemented similarly in both speech and speaker recognition systems. The remaining two steps are implemented differently in

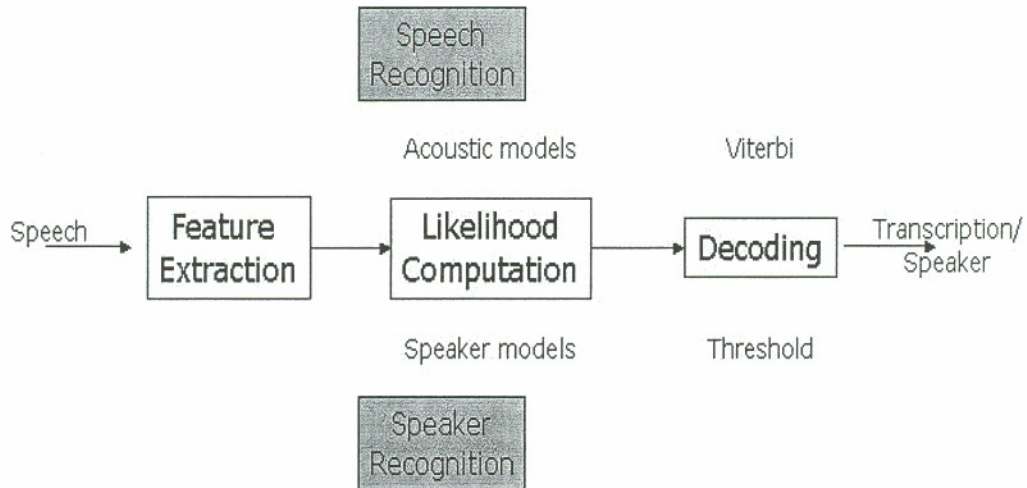


Figure 1.1: Block diagram of speech and speaker recognition systems

these systems.

In the feature extraction step, a set of parameters are estimated from speech signal. This thesis is a study of different type of features, and their usefulness in speech and speaker recognition. Therefore, this step is described in detail in the next section (Section 1.2).

In the likelihood computation step, the likelihood of the features is computed with respect to the speech and speaker models. This is typically performed independently for each feature vector using all the phone or speaker models. The models are trained using the same set of features extracted from a large independent speech corpus.

The type of speech models depends on the task. For a small vocabulary task like digit recognition, speech is modeled as a sequence of words (digits) [55]. For a larger vocabulary tasks, speech is modeled as a sequence of context-dependent or context-independent phones [35]. Each word or phone model is based on Markov model (MM) topology [37]. It has states which correspond to a particular part of the word or the phone, and a transition arcs that correspond to transitions between states. Each state is typically modeled as mixture of Gaussian distributions or it is modeled using a neural network (NN) [22].

In speaker recognition, a speaker can be modeled as a collection of phone models or a single model [19]. Each model can be based on different MM topologies, starting from just one-state to multiple-state, and from left-to-right to fully connected (or ergodic)

models [31]. Each state distribution can be further modeled as a mixture of Gaussians or using a NN. In the speaker recognition research, one-state, multiple Gaussian component model is used most often and it is referred as a Gaussian mixture model (GMM) [12]. This model has been shown to be effective in text-independent speaker recognition. In practice, speaker recognition is performed in text-dependent mode, where the text is a short string of words, such as name, address, account number. This approach is dominated by models based on HMM.

Typically, two different models are used in speaker recognition - speaker independent model or universal background model (UBM) and speaker dependent (SD) model. UBM is trained using large number of speakers, and it does not represent any particular speaker. SD model is trained (or adapted) using the speech from only one speaker.

In the final step, the sequence of likelihoods is analyzed to get the most likely sequence of phones or the most likely speaker given the input speech. In speech recognition, viterbi algorithm is used to obtain this sequence [37]. The search for a sequence is constrained using language models. These models contain multiple pronunciations of words, and probabilities of words and word sequences based on the language. In text-independent speaker recognition, likelihood per speaker is averaged over the complete utterance. The likelihood using SD model is normalized with respect to likelihood using SI model. If the normalized likelihood is greater than a pre-calculated threshold then that speaker is declared as the hypothesized speaker of the utterance. In text-dependent speaker recognition, the most likely sequence of phones is used to obtain the likelihood of the utterance using SD and SI models. The normalized likelihood is used the same way as described before to validate the identity of the speaker.

1.2 Feature Extraction

1.2.1 Feature Characteristics

In this subsection, we present four measures for characterizing a set of features - 1) compactness, 2) correlation, 3) behavior of distribution, 4) relevance for a given task. Compactness measures to number of features in a set. Correlation measures to the relationship

between different features. If a feature is completely predictable by another feature then the correlation between them is 1. If a feature can not be predicted from another feature then the correlation between them is 0. If the features have Gaussian distribution then the uncorrelated features are also independent features. This means that each feature contributes completely new information to the task. Features need to have a well-behaved distribution so that it can be modeled using few training samples and parameters. Relevance refers to the requirement that features must carry information that is relevant for the task.

First three requirements are important for modeling the distribution of features. Features satisfying these requirements need a smaller training dataset, and fewer parameters to model its distribution. These requirements can be measured without any additional information such as class labels. Suitability of the features for a given task can be measured only when the class labels are available, that is, when the task is to classify features into n classes and the training data is labeled by these classes.

1.2.2 Conventional Feature Extraction

In this subsection, we describe the processing steps used to convert a speech signal into a set of commonly used features used for speech recognition. At each step, we analyze output features to measure 4 parameters - 1) dimension, 2) correlation matrix, 3) histogram of a feature element, and 4) F-ratio. Lets assume that X is the feature vector. Dimension of the feature vector shows the compactness. Correlation matrix shows the correlation across different features. Sample histogram of a feature element shows the nature of its distribution assuming that each element is independent. In this work, it is characterized by coefficient of kurtosis (γ_2), and coefficient of skewness (γ_1) [47],

$$\gamma_1 = \frac{\mu_3}{\sigma^3}$$

$$\gamma_2 = \frac{\mu_4}{\sigma^4}$$

where, μ_3 and μ_4 are the third and fourth central moments, and σ is the standard deviation. Kurtosis measures the tails of the distribution and skewness measures the symmetry of the distribution. Values of γ_2 range from $-\infty$ to $+\infty$ where negative values correspond to shorter tails and positive values correspond to longer tails. Values of γ_1 also range from

– inf to +inf where negative values indicate that distribution is skewed to the left and positive values indicate that distribution is skewed to the right. In general a smaller γ_1 and γ_2 correspond to a well-behaved distribution. F-ratio compares across-class covariance (A_c) with within-class covariance (W_c),

$$A_c = \sum_i \frac{N_i}{N} (\bar{X}_i - \bar{X}_{..}) (\bar{X}_i - \bar{X}_{..})^t$$

$$W_c = \frac{1}{N} \sum_i \sum_j (X_{ij} - \bar{X}_{..}) (X_{ij} - \bar{X}_{..})^t$$

where i denotes classes, j denotes the sample index within each class, N_i denotes the number of samples in each class, N denotes the total samples, \bar{X}_i is mean of each class, and $\bar{X}_{..}$ is the mean of the data. F-ratio is defined as $trace(W_c^{-1}A_c)$. This is a measure of separability of the classes. Its significance is measured by degrees of freedom in the numerator ($N_i - 1$) and the denominator ($N - 1$). Higher F-ratio refers to better separability between the classes¹. In this work, sample estimates of these parameters are obtained using a set of 41 phones² and approximately 500000 feature vectors from OGI Stories database [46]. Therefore, the degrees of freedom in the numerator and denominator are 40 and inf, and they are same for all feature-sets.

Speech signal, $s(t)$, is recorded using a microphone, digitized using A/D converter at 8000 Hz, and stored on the disk. The digitized speech is divided into overlapping segments of approximately 20-35 ms (160-280 samples). The adjacent segments overlap approximately 10-15 ms (80-120 samples). This operation is referred to as "windowing" of speech signal. Mathematically,

$$s(p, n) = \sum_n s(t)h(t - pk - n)$$

, where $h(n)$ is a square window of length 160 - 280 samples and p defines the shift across adjacent segments, and n defines the length of the window.

Figure 1.2 shows the result of our analysis using windowed-speech signal, $s(p, n)$ as features. The feature set contains 200 samples. The correlation matrix shows significant correlations between different feature elements. Distribution of the center element of the

¹This is described in detail in Chapter 5

²Phone is smallest unit of speech sound.

feature vector is similar to Laplace distribution with $\gamma_2 = 91.52$ and $\gamma_1 = 0.25$. It is a symmetrical distribution with very long tails. These features have the F-ratio of 0.018, which means that phone classes are highly overlapped in this space.

The speech waveform in each segment is filtered using a high pass filter to remove the DC bias of the microphone and to emphasize the high frequency components. This operation is referred as "pre-emphasis of speech". The waveform is multiplied by Hamming window to minimize the discontinuities at the edges of the window. It is converted to the frequency domain using short-time Fourier transform. Note that the Fourier transform of a real signal is a complex and symmetric signal. The complex signal is converted to a real signal by calculating the magnitude, and only half of it is retained. The sequence of operations results in approximately 129 parameters per segment where the corresponding frequencies are uniformly distributed between 0-4000Hz. The set of these parameters is referred as the "speech spectrum". Mathematically,

$$S(p, w) = \sum_n (s(p, n) * h_{win}(n)) e^{-jwn}$$

, where h_{win} refers to hamming window, $S()$ refers to speech spectrum, and w refers to the discrete frequency values from 0 to π .

Figure 1.3 shows the result of our analysis using speech spectrum, $S(p, w)$, as features. The feature set now contains 129 measurements. The correlation matrix still shows significant correlations across different elements. The features have a one-sided distribution, similar to exponential distribution with $\gamma_2 = 4.13 \times 10^5$ and $\gamma_1 = 50$. It has become a right-skewed distribution with very long tail. The F-ratio using these features is 0.214. Improved F-ratio of the features is a result of non-linear process of calculating the magnitude spectrum.

Human perceptual experiments have shown that human hearing has a non-uniform frequency resolution. To emulate the human processing, previous researchers have proposed warped frequency scales like MEL scale [57, 54] and Bark scale [16, 24]. The speech spectrum is converted to the non-uniform frequency scale as follows. First, triangular or trapezoidal filters are designed on the non-uniform frequency scale. Then, they are mapped to the uniform frequency scale. The speech spectrum is multiplied by each of the

filter coefficients and the resulting values are averaged over the span of the filter. This results in approximately 15-23 point MEL/Bark warped spectrum. Mathematically,

$$\tilde{S}(p, m) = \sum_n S(p, w) FB(w, m)$$

, where m refers to the filter-bank index, and $FB(w, m)$ are m^{th} filter coefficients.

Figure 1.4 shows the result of our analysis using energies from 15 filters on Bark scale, $\tilde{S}(p, m)$, as features. Note that the features set is now reduced to 15 elements. The correlation matrix of the features is similar to the earlier step. The distribution has $\gamma_2 = 788$ and $\gamma_1 = 21$. It is a non-symmetrical, right-skewed distribution. F-ratio for these features is 0.18, which is similar to the one obtained using 129 spectral values. This means that these 15 dimensional feature space has almost the same separation between phones as the 129 dimensional speech spectrum.

In the final processing steps, a non-linear transformation such as log or cube-root is applied on the features and the features are projected on discrete cosine bases. With MEL frequency warping, the resulting features are referred as Mel Frequency Cepstral Coefficients (MFCCs). Mathematically,

$$\begin{aligned} \bar{S}(p, m) &= \log \tilde{S}(p, m) \\ C(p, l) &= \sum_m \bar{S}(p, m) CEP(m, l) \end{aligned}$$

, where l represents the cepstral coefficient, and $CEP()$ is matrix of cepstral transformations.

The results of our analysis of the features after logarithmic transformation, $\bar{S}(p, m)$, and after projection on discrete cosines, $C(p, l)$, are shown in Figures 1.5 and 1.6. Results show that distribution of the features has $\gamma_2 = 2.33$ and $\gamma_1 = 0.098$. It is approximately symmetric and has balanced tails which is close to a Gaussian distribution. The F-ratio of these features increases from 0.18 to 4.22³. However, there is still significant correlation among different elements of the feature set. The projection of features on discrete cosine

³Note that F-ratio of the power spectral energies also improves from 0.18 to 3.6 after logarithmic compression. This shows that Bark frequency warping improves the F-ratio of the features.

Another commonly used compression is cube-root compression. The F-ratio after this compression is 1.38 which is higher than uncompressed spectral energies but lower than logarithmic compression.

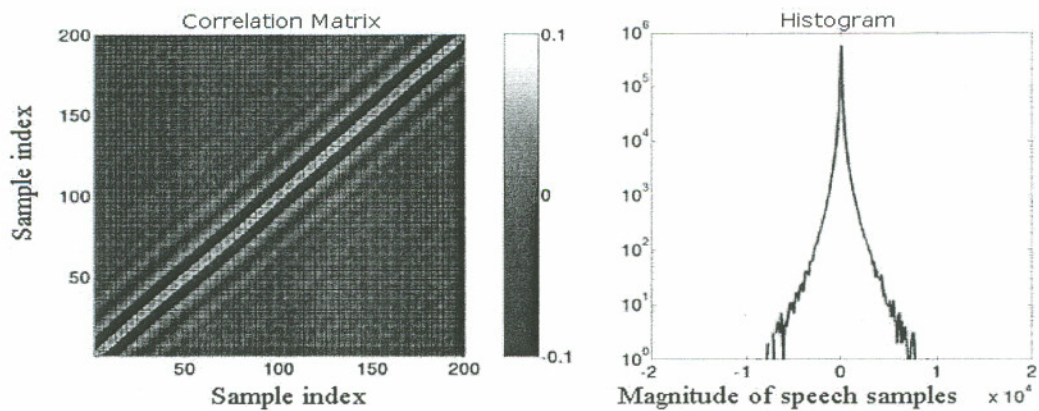


Figure 1.2: Speech samples as features

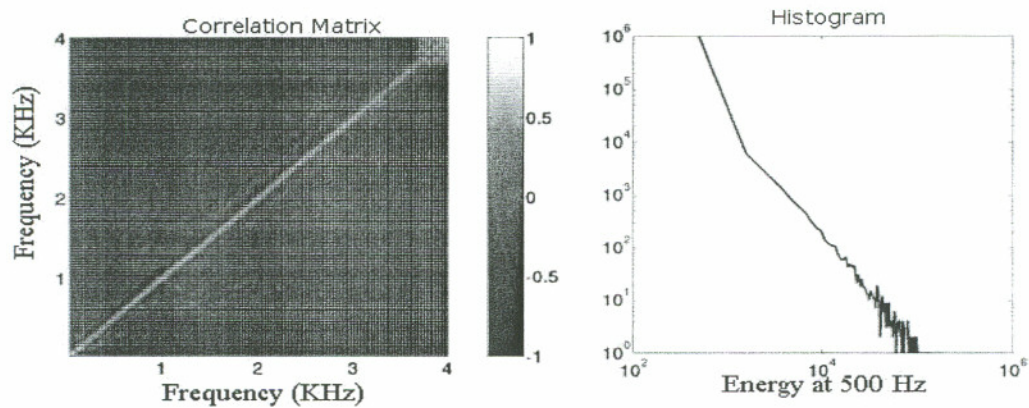


Figure 1.3: Power spectral energies as features

bases results in the best features so far. These features are only 15 dimensional versus 256 dimensional waveform. They have almost diagonal correlation matrix. The histogram of the features shows a very well behaved distribution ($\gamma_2 = 2.59$ and $\gamma_1 = -0.0849$). Since F-ratio is invariant to linear transformation, they have the same F-ratio as the logarithmic energies.

1.3 Variability in Features

As mentioned before, the speech signal, $s(t)$, has variations due to the nature of linguistic message, different speaker characteristics, and different communication channel characteristics. These variations are referred as variability in the speech signal due to different

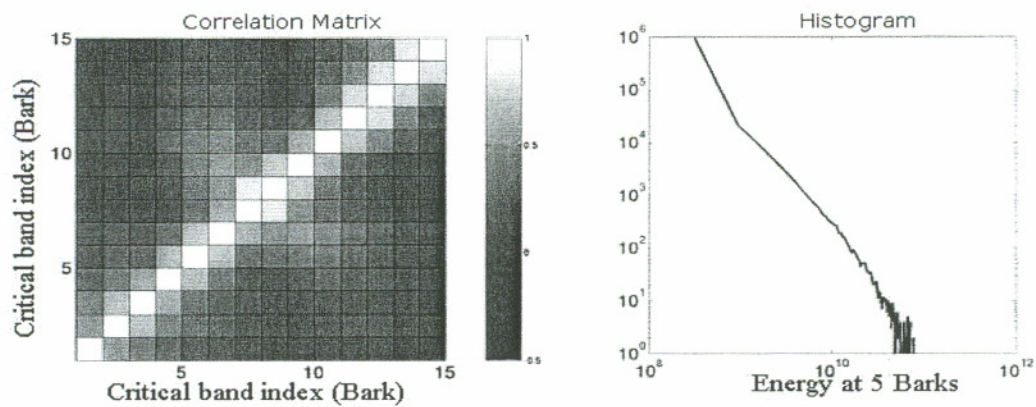


Figure 1.4: Bark-warped power spectral energies as features

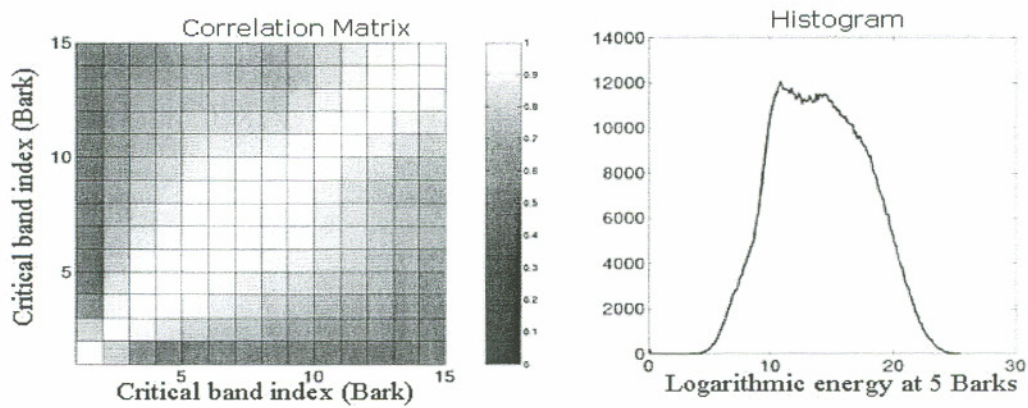


Figure 1.5: Bark-warped power spectral energies with logarithmic compression as features.

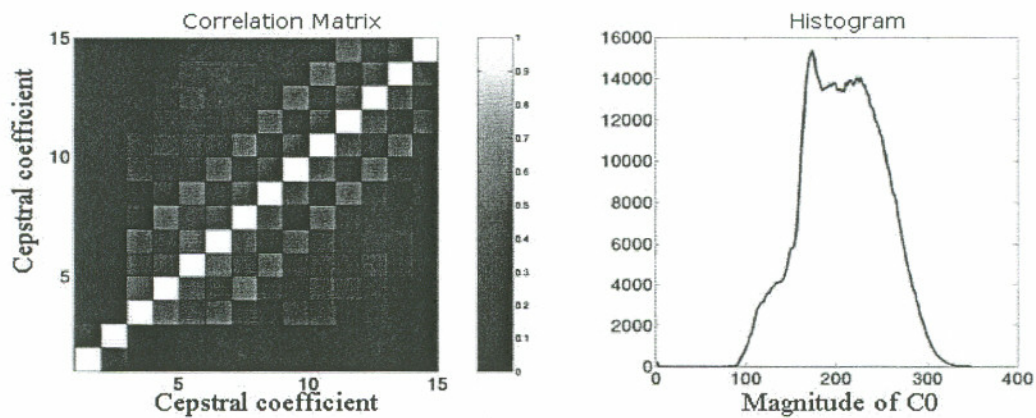


Figure 1.6: Discrete cosine transform coefficients as features

sources. In this thesis, variability in the speech signal is attributed to three sources - 1) language, 2) speaker, and 3) communication channel. Subsequently, total variability in the signal is divided into four types - 1) phone (or inter-phone) variability, 2) speaker variability, 3) channel variability, and 4) residual variability. These types are explained as follows.

1.3.1 Phone (or inter-phone) Variability

Speech signal has variability due to language. The language is used to form a communication message. If the message assumed to a sequence of phones then the variability due to language can be interpreted as variability due to different phones. We refer to this variability as (inter-phone or) phone variability. It is a useful variability in speech recognition.

1.3.2 Speaker Variability

Speech signal has variability due to speakers. A speaker can be described in terms of different characteristics such as vocal tract length, pitch, speaking rate, speaking style, accent. The variability due to speakers is caused by differences in these characteristics. We refer to this variability as speaker variability. It is a useful variability for speaker recognition task.

1.3.3 Channel Variability

Speech has variability due to different communication channels. This is a harmful variability for both speech and speaker recognition.

A communication channel is defined as the path traveled by the speech signal from the speaker's mouth to the listener's ears. It includes communication medium, communication devices, and coders and decoders. Variation in the medium is an environmental noise which is typically modeled as additive distortion. Variation in the communication devices such as telephone handsets, is typically modeled as convolutive distortion. Effect of coders and decoders is modeled as a non-linear distortion in the signal. In this work, channel

variability measures the effect of communication devices like handsets in the telephone. The remaining effects are measured by the residual variability.

1.3.4 Residual Variability

Residual variability measures the variability that is not accounted for the three variabilities described above. This includes variability due to phonetic context, coarticulation and others. Note that its contribution reduces when one of these effects are studied independently.

1.4 Contributions of the Work

The fundamental hypotheses of this work are

1. variability in the feature-set is related to its performance of speech and speaker recognition tasks, and
2. results of analysis of variability can be used to derive new feature-sets that are robust with respect to the mismatch in training and testing.

The hypotheses are examined as follows. First, multivariate analysis of variance (MANOVA) is used to study variability in a feature-set. It is shown that results of MANOVA are related to performance of the features on speech and speaker recognition tasks. Results of MANOVA are used with linear discriminant analysis (LDA) to derive discriminant features for speech recognition task. The discriminant features are shown to improve the performance of speech recognition systems. The contributions of this thesis are explained as follows.

- **Robustness of MANOVA to deviations from normality** - As explained in Chapter 2, MANOVA assumes that features have a Gaussian distribution. In practical situations, the distribution of features deviates from this assumption. We compare the results of MANOVA with another study [30] that uses non-parametric techniques to model the feature distribution. We observe that the results from both

studies are similar, which shows robustness of MANOVA to deviations from normality. Other studies [20, 29, 10] have also observed robustness of this analysis to deviations from normality.

- **Longer-term effects of the phone** - Using MANOVA, we show that variability due to phones lasts for about 250 ms around it. This means that pronunciations of current phone affects the nature of at least the neighboring phones assuming that an average length of a phone is approximately 80 ms.
- **Relative phone variability and speech recognition performance** - Results of MANOVA can be used to compare different feature representations. In this work, we compare different types of features such as features with and without delta features; features with and without commonly used feature transformations like RASTA filtering, features with and without mean and variance normalization; features before and after concatenating TRAPS [25] features. In all these comparisons, we show that performances of the resulting features on speech recognition tasks is related to the relative contribution of phone variability.
- **Relative speaker variability and speaker recognition performance** - We compare speaker variability in the features for different broad phonetic categories. This is compared to the speaker recognition performance obtained using these categories. Results show that relative speaker variability in the features is related to the speaker recognition performance.
- **Relationship between MANOVA and LDA** - MANOVA and LDA are compared. It is observed that both analysis make same assumption about the data. It is shown the results of MANOVA can be grouped into useful and harmful variabilities. These variabilities can be used as an estimates of across-class and within-class covariances. Thus, it is shown that covariances estimated from MANOVA can be used as an input to LDA.
- **Application of results of MANOVA in speech recognition** - Joint time-frequency discriminants are designed using a longer temporal span. It based on

the result that variability due to phones lasts beyond their length. Discriminants are obtained using LDA. Phone variability is used as across-class covariance and remaining variability is used as within-class covariance. It is shown that the discriminant features using 1000 ms of temporal span give significant improvements in the speech recognition performance over the conventional features that use only 90 ms of temporal span.

1.5 Organization of Thesis

Chapter 2 describes MANOVA and uses it to measure the variability in the features. The total variability is analyzed in spectral and temporal domains. In each domain, it is divided into four types as described in Section 1.3. The chapter proposes a method to quantify the contribution of sources in each domain. We compare the results of this analysis with results of two previous analysis. The results show that analysis of variance can be also viewed as analysis of information in speech.

Chapter 3 shows that relative contributions of different variabilities is related to the performance of the features on speech and speaker recognition tasks. We calculate relative speaker variability in the features for broad phonetic categories. This is compared to the speaker recognition performance obtained using those categories. It is shown that the phone categories with the highest relative speaker variability perform the best on the speaker recognition task.

Relative phone variabilities in different feature-sets are compared with their performance on speech recognition task. In first experiment, effect of delta and double-delta features is compared using relative phone variability and speech recognition performance. In second experiment, relative phone variability is measured after different feature transformations, such as RelAtive SpecTrAl (RASTA) [27] filtering, mean and variance normalization, and TempoRAI PatternS (TRAPS) [25] processing. We show that these transformation improve the relative phone variability and also improve the speech recognition performance.

Chapter 4 relates MANOVA with LDA. Both analysis make similar assumptions about

the data. For a given task, results of MANOVA can be grouped into two types of variabilities - useful and harmful variability. They can be used as estimates of across-class and within-class covariances required by LDA. Thus it is shown that results of MANOVA can be used with LDA to design a robust feature-set.

Chapter 5 applies one of the important results from chapter 2 for designing a new set of features for speech recognition. The result shows that effect of current phone lasts beyond its boundaries, and well into the neighboring phones. It is used to derive joint time-frequency discriminants for speech recognition using a longer temporal span. The discriminants are obtained using two methods - joint analysis and combined analysis. It is shown that incorporation of longer time span improves speech recognition performance. It is also shown that combined analysis assumes that time and frequency domains are independent. It requires less training data and generalizes better than joint analysis.

Chapter 6 summarizes the work and suggests future directions.

Chapter 2

Analysis of Variability in Speech

In this chapter, we decompose the variability in speech into four types: linguistic variability, speaker variability, channel variability, and residual variability. Linguistic variability is due to variation across phones in the language, speaker variability is because of the variations in different speaker characteristics such as accent and speaking rate, and channel variability is attributed to the variations in the communication channels, i.e., handset and the telephone line. Residual variability is not associated with any particular source because it shows the effect of all the unaccounted sources such as phonetic context and coarticulation. Apart from the source variabilities, we also study the source dependencies, which are called interaction variabilities.

The chapter is organized as follows. Section 2.1 describes analysis of variance in general. Section 2.2 describes how MANOVA is used to analyze variability in speech. It also describes how the source and interaction covariances are computed from the data. Section 2.3 validates our assumptions behind MANOVA. This is followed by the description of the databases and features in Section 2.4. Sections 2.5 and 2.6 describe variability in spectral and temporal domains. These results are quantified in Section 2.7. Section 2.8 compares our results to the results from the previous studies. It is shown that the analysis of variance in speech is similar to analysis of information in speech.

2.1 ANOVA

Analysis of variance (ANOVA) is used to measure the variation in the data with respect to one factor, for example, the variation in the oil price, X , across different months of year.

This is a general statistical analysis technique which is used in many different areas from psychology to marketing research for product testing. The main idea is to use following model to explain the variation in the data,

$$X_{ij} = \mu + \mu_i + \epsilon_{ij}$$

, where μ is the mean of the data, i is the class index within one factor (there are n classes), μ_i ¹ is the class mean, and ϵ_{ij} is the error in the approximation. This can be also seen as test of equality of means. Given different means, $\mu_1, \mu_2, \dots, \mu_n$, with a common variance σ^2 , we wish to test equality of these means, $\mu_1 = \mu_2 = \dots = \mu_n = \mu$. This is referred to as null-hypothesis H_0 . Using our example, we can say that μ_i is the mean price of oil in a month i , μ is the mean of the oil price over an year, and σ^2 is the variation in the oil price within a month (which is assumed to be same for all the months). We wish to know if the model is good for predicting the variations in the oil prices or if the oil prices across different months are same or not. In other words, is there any significant variation among the average monthly oil prices? How much variation in the data can be explained by average oil price per month?

In order to test the hypothesis, we take a random samples $\{X_{ij}\}$ from the distribution of oil prices per month, $\mathcal{N}(\mu_i, \sigma^2)$, where i denotes month and j denotes the sample number from that month. Let n_i be the samples from each month and $n = \sum_i n_i$, be total number of samples.

$$\begin{aligned}\bar{X}_{..} &= \frac{1}{n} \sum_{i=1}^m \sum_{j=1}^{n_i} X_{ij} \\ \bar{X}_{i.} &= \frac{1}{n_i} \sum_{j=1}^{n_i} X_{ij}\end{aligned}$$

where the dot in the notation shows the index over which the variable is averaged. The total variance of the data is decomposed into two terms.

$$\begin{aligned}\frac{1}{n} \sum_{i=1}^m \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_{..})^2 &= \frac{1}{n} \sum_{i=1}^m (\bar{X}_{i.} - \bar{X}_{..})^2 + \frac{1}{n} \sum_{i=1}^m \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_{i.})^2 \\ \sigma_T^2 &= \sigma_F^2 + \sigma_E^2\end{aligned}$$

¹bold symbol is used for true estimate and normal symbols are used for sample estimates

where σ_T^2 is the estimate of total variance, σ_F^2 is the estimate of the variance across different months, and σ_E^2 is the estimate of the variance within each month. Note that $\frac{1}{n} \sum_{i=1}^m \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_{..})^2$ is a biased estimated of the total covariance. We assume that $n \rightarrow \text{inf}$ and the bias is not significant.

If the null-hypothesis is true then, σ_E^2 is approximately equal to σ^2 , or $\frac{\sigma^2}{\sigma_E^2}$ is very close to 1. If the means, μ_i , are different then the ratio becomes greater than 1. The ratio is called as F-ratio, $F = \frac{\sigma^2}{\sigma_E^2}$. It has F distribution with $m - 1$ and $n - m$ degrees of freedom. The tests of significance is performed by comparing the empirical F with $F_\alpha(m - 1, n - m)$. Here α is called as significance value. If $F \geq F_\alpha(m - 1, n - m)$, then H_0 is rejected at the significance α .

ANOVA makes following assumptions about the data - 1) normality, 2) homogeneity of variances, and 3) additive model. Normality refers to the assumption that the complete data and data within each factor is Gaussian distributed. Homogeneity refers to the assumption that variances within each factor are same. Third assumption is related to the fact that the variances are assumed to be additive.

2.2 MANOVA

Multivariate analysis of variance (MANOVA) is used to measure the variation in the data, $\{X\}$, with respect to more than two factors. In this work, we use three factors - phone, speaker and channel. They correspond to the sources of variability in speech. The underline model of MANOVA is

$$X_{ijkl} = \bar{X}_{....} + \bar{X}_{i...} + \bar{X}_{ij..} + \bar{X}_{ijk.} + \epsilon_{ijkl} \quad (2.1)$$

where, $i = 1, \dots, p$ represents phones; $j = 1, \dots, s$ represents speakers; and $k = 1, \dots, c$ represents channels. This equation shows that any feature vector, X_{ijkl} , can be approximated using a sum of $\bar{X}_{....}$, the mean of the data; $\bar{X}_{i...}$, mean of the data for the phone i ; $\bar{X}_{ij..}$, mean of the data for speaker j and phone i ; $\bar{X}_{ijk.}$, mean of the data for speaker j , phone i and channel k ; and ϵ_{ijkl} , an error in this approximation. Using this model, the total covariance can be decomposed as follows

$$\Sigma_{total} = \Sigma_{phone} + \Sigma_{speaker} + \Sigma_{channel} + \Sigma_{residual}$$

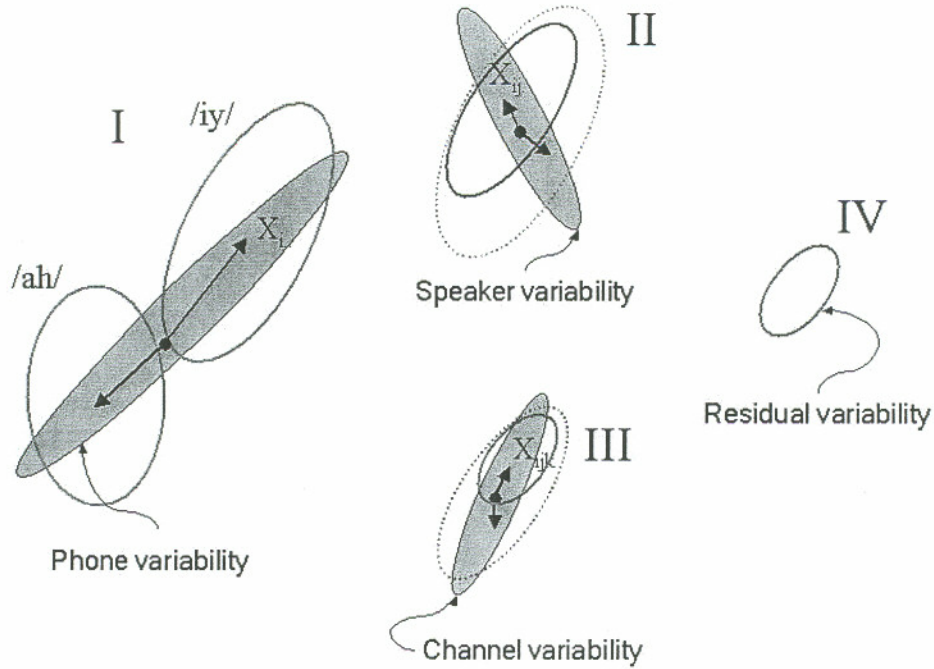


Figure 2.1: Sources of variability in speech using MANOVA

where

$$\begin{aligned}\Sigma_{phone} &= \sum_i \frac{N_i}{N} (\bar{X}_{i\dots} - \bar{X}_{\dots})^t (\bar{X}_{i\dots} - \bar{X}_{\dots}) \\ \Sigma_{speaker} &= \sum_i \sum_j \frac{N_{ij}}{N} (\bar{X}_{ij..} - \bar{X}_{i\dots})^t (\bar{X}_{ij..} - \bar{X}_{i\dots}) \\ \Sigma_{channel} &= \sum_i \sum_j \sum_k \frac{N_{ijk}}{N} (\bar{X}_{ijk.} - \bar{X}_{ij..})^t (\bar{X}_{ijk.} - \bar{X}_{ij..}) \\ \Sigma_{residual} &= \frac{1}{N} \sum_i \sum_j \sum_k \sum_l (X_{ijkl} - \bar{X}_{ijk.})^t (X_{ijkl} - \bar{X}_{ijk.})\end{aligned}$$

and, N is the data size and N_{ijk} refers to the number of samples associated with the particular combination of factors (indicated by the subscript).

Assuming that each feature vector is labeled using the phone, speaker and channel, the covariance term are computed as follows. First, all the feature vectors (X) belonging to each phone i are collected and their mean ($\bar{X}_{i\dots}$) is computed. The covariance of these phone means is called the phone variability (shown as the shaded oval in Figure 2.1

(I), Σ_{phone} . Next, the data for each speaker j within each phone i is collected and the mean of the data ($\bar{X}_{ij..}$) is computed. The covariance of the means of different speakers (shown as the shaded oval in Figure 2.1 (II)) averaged over all phones is called the speaker variability ($\Sigma_{speaker}$). Similarly data for each channel k within a phone i and a speaker k is used to compute the mean of the channel ($\bar{X}_{ijk.}$). The covariance of the channel means (shown as the shaded oval in Figure 2.1 (III)) averaged over all phones and all speakers is termed as the channel variability ($\Sigma_{channel}$). All the variability in the data is not explained using these sources. The unaccounted sources, such as context and coarticulation, cause variability in the data collected from one speaker speaking one phone through one channel (Figure 2.1 (IV)). The covariance within each phone, speaker, and channel is averaged over all phones, speakers, and channels, and the resulting covariance is called residual variability ($\Sigma_{residual}$).

If we denote $OP(y) = y^t y$, then $\Sigma_{speaker}$ and $\Sigma_{channel}$ can be further expanded to

$$\sum_i \sum_j N_{ij} OP(\bar{X}_{ij..} - \bar{X}_{i...}) = \sum_j N_j OP(\bar{X}_{.j..} - \bar{X}_{....}) + \sum_i \sum_j N_{ij} OP(\bar{X}_{ij..} - \bar{X}_{i...} - \bar{X}_{.j..} + \bar{X}_{....}) \quad (2.2)$$

$$\begin{aligned} \sum_i \sum_j \sum_k N_{ijk} OP(\bar{X}_{ijk.} - \bar{X}_{ij..}) &= \sum_k N_k OP(\bar{X}_{..k.} - \bar{X}_{....}) \\ &+ \sum_i \sum_k N_{ij} OP(\bar{X}_{i.k.} - \bar{X}_{i...} - \bar{X}_{..k.} + \bar{X}_{....}) \\ &+ \sum_j \sum_k N_{jk} OP(\bar{X}_{.jk.} - \bar{X}_{.j..} - \bar{X}_{..k.} + \bar{X}_{....}) \\ &+ \sum_i \sum_j \sum_k N_{ijk} OP(\bar{X}_{ijk.} - \bar{X}_{ij..} - \bar{X}_{i.k.} \\ &- \bar{X}_{.jk.} + \bar{X}_{i...} + \bar{X}_{.j..} + \bar{X}_{..k.} - \bar{X}_{....}) \quad (2.3) \end{aligned}$$

In equation 2.2, the first term is the covariance of the speaker means ($\bar{X}_{.j..}$). it is referred as the global speaker variability. The second term is the interaction between phone and speaker factors, which is called phone-specific speaker variability. Similarly, in equation 2.3, the first term is the covariance of the global channel mean ($\bar{X}_{..k.}$). It is referred as the global channel variability. The second term is the interaction between channel and phone factors and it is called phone-specific channel variability. The third

term is the interaction between channel and speaker factors which is referred as speaker-specific channel variability. The last term in equation 2.3 is the interaction among all the three factors. It is referred as phone- and speaker-specific channel variability.

The interaction terms are important for following reason. The most commonly used preprocessing technique in speech and speaker recognition is called utterance-based mean subtraction (UMS)²[38]. It is used to remove the global speaker and channel variabilities. The effect of UMS can be modeled in MANOVA by setting the first terms from equations 2.2 and 2.3 to zero. This shows that the remaining speaker and channel variabilities are only due to their interactions. In the remaining sections, we will present results of MANOVA with and without UMS to show the contribution of interaction variabilities in speech.

2.3 Assumptions of MANOVA

As mentioned in 2.1, MANOVA makes three assumptions about the features. First two assumptions are about the normality of the data and homogeneity of distributions within different factors. Although this may not be valid under practical conditions, it has been shown that results of MANOVA are robust to deviation from normality [20]. We have also shown the robustness of MANOVA by comparing its results with results of a study that did not make any assumptions about the distribution of features.

Third, the variability in features can be decomposed using additive model. Use of this model is justified by assuming a source-filter model of speech production, and by assuming that channel can be modeled as time-varying filter as follows. Lets assume that $r(n)$ is the excitation, $f(n)$ is the vocal tract filter, and $c(n)$ is the channel filter. So, resulting speech signal $s(n)$ can written as,

$$s(n) = e(n) * f(n) * c(n),$$

where $*$ is convolution operator. Further, we assume that vocal tract filter can be decomposed into two different filters, filter corresponding to the phone, $p(n)$, and filter

²It is also known as cepstral mean subtraction (CMS) when the operation is done in “cepstral” domain

corresponding to the speaker, $sp(n)$. So, we can rewrite the above equation as,

$$s(n) = e(n) * p(n) * sp(n) * c(n).$$

If we transform the signal using Fourier transform, the convolutive effect becomes multiplicative,

$$S(f) = E(f) \times P(f) \times SP(f) \times C(f).$$

Further, if we compress $S(f)$ using logarithmic non-linearity then the multiplicative effect becomes additive,

$$\log S(f) = \log E(f) + \log P(f) + \log SP(f) + \log C(f).$$

In this work, we used logarithmic filter-bank energies as features for MANOVA where the effect of different sources is approximately additive.

The proposed nesting of the sources, phone \rightarrow speaker \rightarrow channel, is only one way of analyzing variability in speech. The analysis can be performed using a different hierarchy where speaker or channel is used as the first source. This will change the results of MANOVA. Note that any hierarchy of the sources will not change their global effects (Σ_p and first terms in Equations 2.2 and 2.3) and their interactions (second term in Equation 2.2, and second, third and fourth terms in 2.3). The difference in the results will be due to difference in the combination of source and interaction effects. For example, in the proposed nesting of phone \rightarrow speaker \rightarrow channel, the interaction between phones and speakers is included in speaker variability. If the sources are nested as speaker \rightarrow phone \rightarrow channel, then the interaction between phones and speakers will be included in the phone variability.

Another issue in deciding the hierarchy is the definition of speaker and channel variability. As mentioned before, speaker variability can be defined as the sum of two terms - global speaker variability and phone-specific speaker variability. Global speaker variability is the covariance of means of the utterances from different speakers. In speech and speaker recognition, this variability is commonly removed using transformation like UMS. The remaining speaker variability is due to interaction between phones and speakers. We are interested in these interactions, specifically, for their applications in speaker recognition.

The same explanation applies to channel variability too. Since global channel variability is also commonly removed using UMS, we are interested in the remaining channel variability. This is due to interactions of channels with phones and speakers. To summarize the discussion, we are interested in global phone variability and interaction variability between phones and other factors. Therefore the hierarchy, phone \rightarrow speaker \rightarrow channel, is used in this work.

Finally, it is noted that MANOVA makes some assumption (as described above) about the data and, in this application, these assumptions are not completely satisfied. However, we will show that statistics estimated using results of MANOVA are related to the performance of features on speech and speaker recognition task. Further, we will also show that results of MANOVA can be used to derive robust features for speech and speaker recognition.

2.4 Database and Features

Three databases are used in this work: HTIMIT, OGI Stories and TIMIT. HTIMIT [13] database is labeled by all the three sources of variability: phone, speaker and channel and is used to obtain the primary results. OGI Stories database is labeled by two sources of variability: phone and speaker+channel. It is the largest database among the three and is used in the subsequent chapters to derive discriminants for speech recognition. TIMIT database is labeled by two sources of variability: phone and speaker. It is used to study the amount of speaker and phone variability in absence of channel variations. It is also a good reference for comparing the results from HTIMIT (it was created by passing a subset of TIMIT utterances through different handsets).

As mentioned before, HTIMIT database is used for analyzing the nature of variability in spectral and temporal domains. Appendix A are thirty-five phones used in this analysis. There are 137 speakers and 8 handsets – 4 electret and 4 carbon-button – used in the following analysis.

Figure 2.2 shows a time-frequency representation of parameters estimated from a

speech signal. The y-axis represents frequency, x-axis represents time, and the darkness of each element shows the energy at a given frequency and time. A spectral vector is defined by the number of points on the y-axis, $S(w, t_m)$. In this work, this vector contains 15 points on Bark spectrum[52]. The vector is estimated at every 10 ms using a 25 ms speech segment. Each vector is labeled by the phone, speaker, and channel label of the corresponding speech segment. A temporal vector is defined by a sequence of points along time at a given frequency, $S(w_n, t)$. In this work, this vector consists of 50 points each in the past and the future with respect to the current observation and the observation itself. As the spectral vectors are computed every 10 ms, the temporal vector represents 1 sec of temporal information. The temporal vectors are labeled by the phone, speaker, and channel label of the current speech segment.

Figure 2.3 shows the difference in the computation of variability in spectral and temporal domain. In spectral analysis (a), 15 dimensional feature is used and it is labeled by the phone, speaker and channel label of the corresponding speech segment. All the vectors labeled as the same phone are collected and the covariance of the phone means is computed the estimate of phone variability in spectral domain. The phone variability in spectral domain refers to the variance of each spectral measurement due to difference in the phones.

The temporal analysis (b) is performed independently for each band using 101 dimensional feature is used. The vector contains current observation, 50 observations in the past, and 50 observations in the future. Since the spectral vectors are estimated at 10 ms, current observation refers to 0 ms, and 50 observations in the past and the future correspond to 500 ms of temporal context. Each vector is labeled by the phone, speaker and channel of the current observation. All the temporal vectors labeled as the same phone are collected and the covariance of the phone means is computed as an estimate of phone variability. The phone variability in temporal domain refers to the variance of each temporal element due to difference in the phones at the current observation. This shows the effect of variation in the phone at the current observation on the temporal observations in the past and the future.

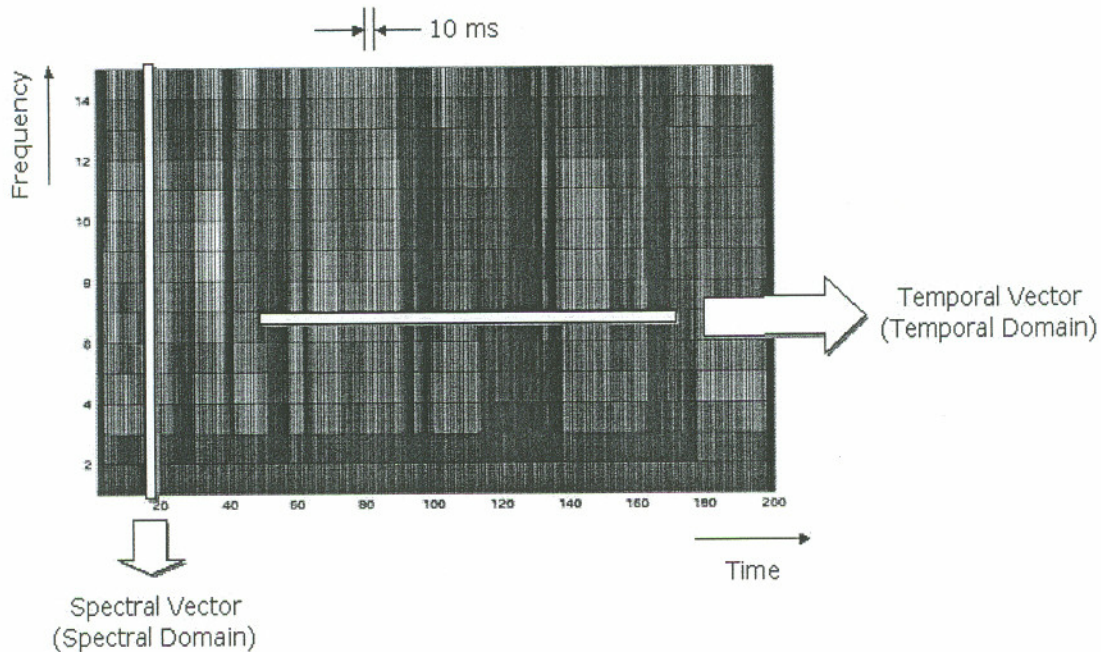
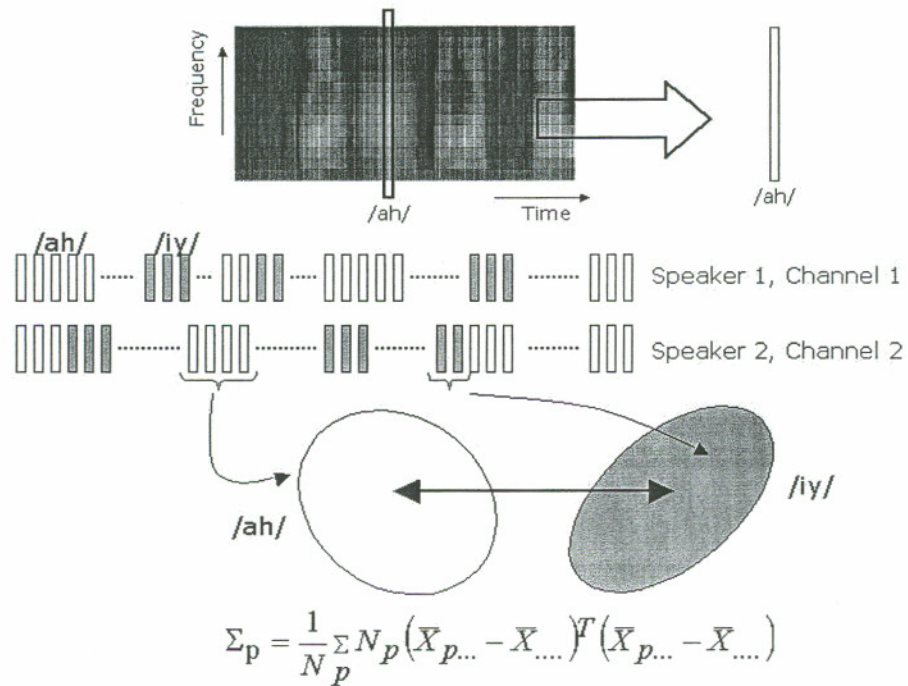


Figure 2.2: Spectral and temporal feature vectors used for MANOVA

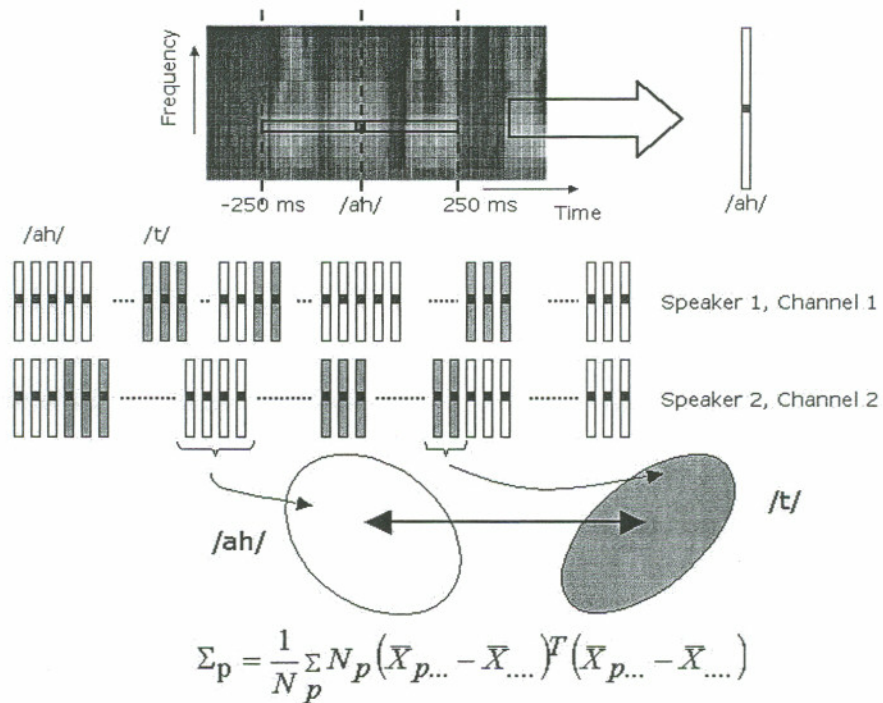
2.5 Nature of Variability in Spectral Domain

Figure 2.4 shows the results of MANOVA in spectral domain. It is assumed that features in spectral domain are independent. Results in spectral domain show variability in different frequency bands because of different phones, speakers, and channels. We make following observations from these results.

1. The region around 5-6 barks (500-600 Hz) contains the highest phone variability. This variability is attributed to characteristics of voiced/unvoiced phones in this region [26] where the voiced phones have high energy in this region and the unvoiced phones have low energy.
2. Without UMS, the channel variability is higher than the speaker variability. This shows that the long-term average of features in time is affected more by the channel than by the speaker. After UMS, however, the speaker variability is higher than the channel variability. Since the speaker and channel variabilities after UMS represent interaction between phone and these sources, it is concluded that different phones contain different amounts of speaker variability. After UMS, the channel variability



(a) Spectral domain



(b) Temporal domain

Figure 2.3: Difference in the computation of phone variability in spectral and temporal domains

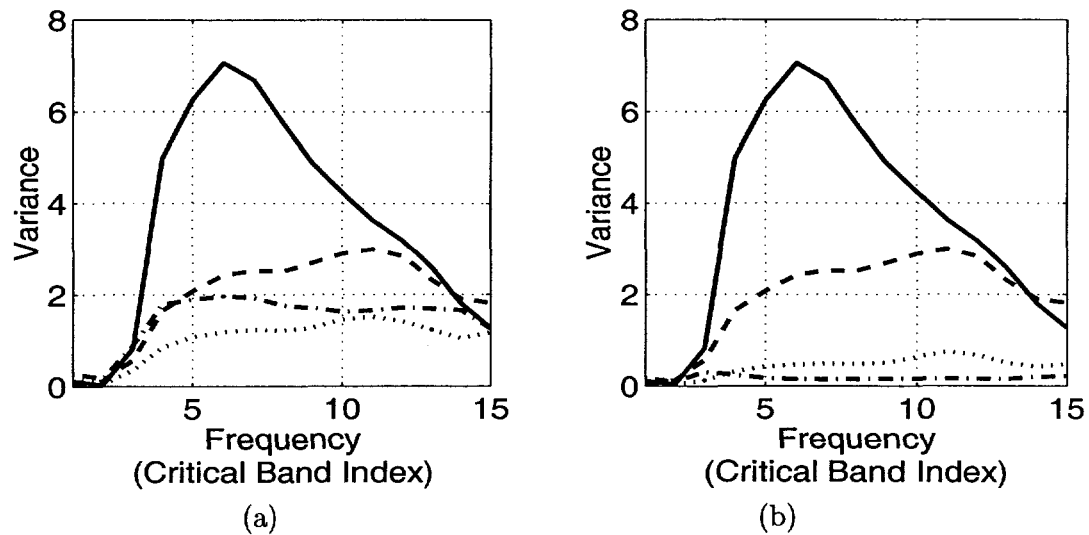


Figure 2.4: MANOVA in spectral domain using HTIMIT database. Before utterance-based mean subtraction (MS) (a) and after UMS (b). Phone information (solid line) (I), speaker information (dotted line), channel information (dash-dot line), and error information (dash line). Note the reduction in speaker and channel information after UMS.

is higher in lower bands whereas the speaker variability is higher in higher bands.

- Note the similarity between the structure of residual variability and the speaker variability. The similarity is due to the fact that different speakers within each phone also differ in the phonetic context in addition to speaker characteristics. This results in interaction between speaker and residual variability. This effect can be minimized by having more speech from different speakers so that the means of different speakers speaking a phone are less sensitive to the variations due to phonetic contexts.

2.6 Nature of Variability in Temporal Domain

In temporal domain, the variability in 15 frequency bands is analyzed independently. However the results from only the fifth band are presented here³ (figures 2.5a and 2.5b). Results from other bands are shown in Appendix C. It is assumed that features in temporal domain are independent. Results in temporal domain show the variability in features in

³The other bands show a similar structure as the fifth band and the conclusions from analysis of the fifth band are applicable to the other bands.

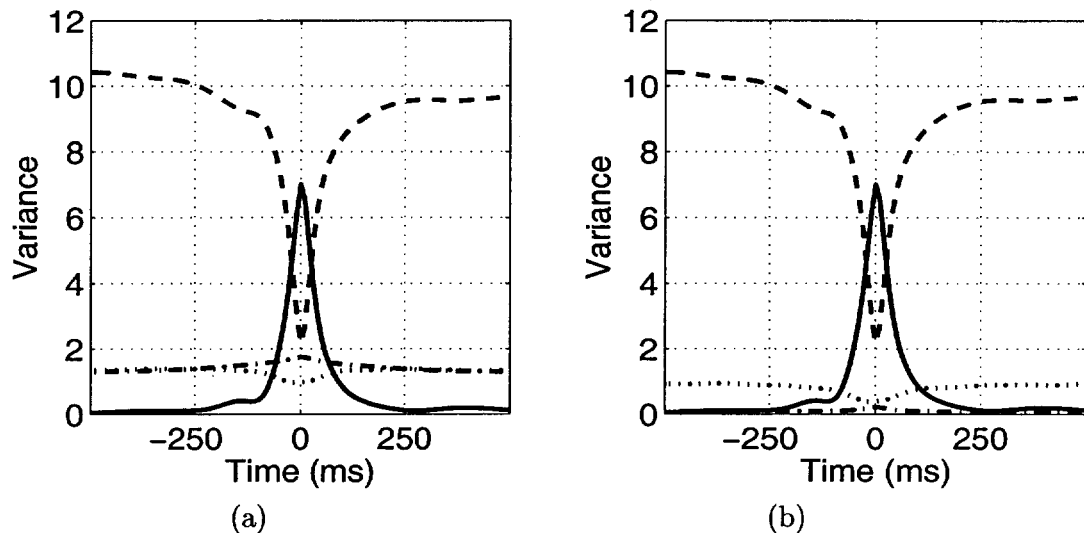


Figure 2.5: MANOVA in temporal domain using HTIMIT database. Before utterance-based mean subtraction (MS) (a) and after UMS (b). Phone variability (solid line), speaker variability (dotted line), channel variability (dash-dot line), and error variability (dash line). Note the reduction in the speaker and channel variability after UMS.

the past and the future due to the variation in phones, speakers, and channels at the current time instance. Following observations are made from these results:

1. The phone variability is highest at the center ($t = 0$) and it drops to a small value 250 ms away from the center. This shows that the variability introduced by phones at the current frame lasts for approximately 250 ms in the past and in the future. This conclusion forms the basis for designing long filters (~ 1 sec) in the temporal domain.
2. The speaker and channel variability before and after UMS is approximately constant across time. After UMS the speaker variability is more dominant than the corresponding channel variability.
3. The residual variability is minimum at $t = 0$. Note that it is computed using only the features within a phone. After accounting for speaker and channel variations, this point represents the position-within-phone variability⁴. Beyond the center residual

⁴All the frames labeled by a phone are used for estimating the statistics for $t = 0$

variability increases rapidly and asymptotes to a constant value due to the high variability in the phonetic context.

4. Note the similarity in the structure of the speaker and residual variability around the current frame ($t = 0$). This is due to the fact the different speakers within a phone also differ in their phonetic contexts. This effect is also seen in the spectral domain.

2.7 Contribution of Sources of Variability

After describing the variabilities in HTIMIT, we quantify the contribution of the sources in spectral and temporal domains. Contribution of each source in spectral domain is defined as

$$\text{contribution of source} = \text{trace}(\Sigma_{\text{source}}) / \text{trace}(\Sigma_{\text{total}}). \quad (2.4)$$

This measure assumes that each source and the total variabilities have a joint Gaussian distribution in the feature space⁵. It is independent of scaling and rotation or linear transformation. This measure is also commutative. Therefore, contribution of all sources, their interactions and the residual variabilities sums to 1. Note that *trace* of the covariance is also sum of its eigen values. Therefore the measure shows the relative energy in each source distribution compared with the total energy in the features.

Note that features used in spectral and temporal domain have different dimensionality. Features in spectral domain have 15 dimensions, and those in temporal domain have 101 dimensions. To compare contribution of sources from both domains, we select 15 dimensions from temporal domain using principal component analysis. We compute eigenvectors of Σ_{total} , and select 15 eigenvectors that have the largest eigenvalues (E). The source and total covariance is projected on these dimensions. The resulting covariances are used in equation 2.4 to measure contribution of sources in temporal domain. This is shown as $\text{trace}(E^t \Sigma_{\text{source}} E) / \text{trace}(E^t \Sigma_{\text{total}} E)$.

Table 2.1 (a) and (b) show the contribution of the sources in HTIMIT measured in

⁵In general, it is accurate up to second-order statistics for any kind of distribution

spectral and temporal domains. In each domain, the results are presented before and after UMS. The variance of the estimated source variance is computed using 7 different data-sets from HTIMIT and is indicated in the bracket besides the contribution. Results show that the results are independent of the variations in the data-sets.

Comparison of results across spectral and temporal domains shows that there is less phone variability in temporal domain than spectral domain. This means that phone recognition performance using a time-trajectory from one band will be worse than using the vector of spectral energies. This can be verified using continuous numbers recognition experiments. We choose 8 features from both spectral domain and from temporal domains. In spectral domain, fifteen spectral energies are projected on 8 leading spectral linear discriminants. In temporal domain, 101-point temporal trajectories of spectral energies from fifth critical band are projected on 8 leading temporal linear discriminants. The rest of the setup is identical for both the feature-sets⁶. The recognition results show that spectral feature give 15% word-error-rate and temporal features give 79% word-error-rate. Note that the phone variability in temporal domain is small but it is significant because recognition accuracy using temporal features is better than chance (9%).

Higher residual variability in temporal domain is due to the fact that it also includes the variation in the neighboring phones. It is shown in Appendix B that the contribution of residual variability decreases by modeling the variability due to phonetic context explicitly. The results indicate that using context-dependent phones as units, the residual variability reduces from 82.6% to 56.4%. This also shows that approximately 26% of the variability is due to the neighboring phones. The residual variability can be further reduced by modeling the longer phonetic context in the design of MANOVA.

Table 2.2 and 2.3 compares the contribution of the sources across different databases: HTIMIT, TIMIT and OGI Stories. TIMIT and OGI Stories database are analyzed using different sources than HTIMIT. Variability in TIMIT is analyzed using phone and speaker; and variability in OGI Stories is analyzed using phone and speaker+channel.

The comparison between results from TIMIT and HTIMIT in spectral domain before

⁶It is described in detail in Chapter 5.

UMS (Table 2.2 (a)) shows that adding channel variability reduces the contribution of the phone and the residual variability. Speaker variability is almost same across these databases. Similar results from OGI Stories show that the phone variability is reduced further by adding the telephone line variations along with the handset variations. After UMS (Table 2.2 (b)), the source contributions between OGI Stories and HTIMIT become comparable. The differences in the speaker and channel interaction variability and the residual variability can be because HTIMIT consists of read speech and it has artificial channel variations.

Similar observations can be made in temporal domain before UMS (Table 2.3 (a)). After UMS (Table 2.3 (b)), the source interactions in HTIMIT and OGI Stories is almost similar. The difference in the phone and residual variabilities is due to lack of sufficient context in estimating the phone variability. This makes the phone variability sensitive to the unaccounted sources which increases its contribution and decreases the contribution of the residual variability.

Note that the proposed measure is not the only way to quantify the contribution of different sources. Appendix E describes other measures that can be also used to describe source contributions. Some of these measures are also used in subsequent chapters to show correlation of results of MANOVA with speech recognition performance.

2.8 Results in Perspective

This section compares the results from this study to the results from two previous studies [15, 30]. First study [15] uses hierarchical ANOVA to decompose the variability in TIMIT database. The total variability is decomposed using nine factors. We map these factors to the three factors from our study and show that results from the two studies are similar. This study measures variability as variance or using second-order statistics. Second study [30] showed that features do not have a Gaussian distribution. Therefore, variabilities should be computed using higher-order statistics. The study used mutual information (MI) to characterize the variability in speech and the information is computed using non-parametric density estimation techniques. We convert results of MANOVA to mutual

Table 2.1: Contribution of sources in spectral domain (a) and temporal domain (b) in HTIMIT database. Numbers in the bracket indicate the variation in the results over 7 different data-sets.

| (a) Spectral domain | | |
|---------------------|---------------------|---------------------|
| source | % contribution | |
| | Before UMS | After UMS |
| phone | 44.8 (± 0.35) | 58.1 (± 0.45) |
| speaker | 12.8 (± 0.16) | 6.9 (± 0.19) |
| channel | 18.5 (± 0.58) | 2.9 (± 0.27) |
| residual | 24.7 (± 0.60) | 31.9 (± 0.50) |

| (b) Temporal domain | | |
|---------------------|---------------------|---------------------|
| source | % contribution | |
| | Before UMS | After UMS |
| phone | 6.1 (± 0.03) | 7.0 (± 0.03) |
| speaker | 11.9 (± 0.11) | 7.2 (± 0.12) |
| channel | 11.5 (± 0.24) | 0.8 (± 0.10) |
| residual | 70.2 (± 0.17) | 82.6 (± 0.25) |

Table 2.2: % contribution of sources in spectral domain without (a) and with UMS (b)

| (a) Without UMS | | | |
|-----------------|----------------|-------------|-------|
| source | % contribution | | |
| | HTIMIT | OGI Stories | TIMIT |
| phone | 44.8 | 35.3 | 57.8 |
| speaker | 12.8 | 41.1 | 13.1 |
| channel | 18.5 | | x |
| residual | 24.7 | 23.5 | 28.9 |

| (b) With UMS | | | |
|--------------|----------------|-------------|-------|
| source | % contribution | | |
| | HTIMIT | OGI Stories | TIMIT |
| phone | 58.1 | 56.3 | 62.7 |
| speaker | 6.9 | 7.4 | 5.9 |
| channel | 2.9 | | x |
| residual | 31.9 | 36.3 | 31.4 |

Table 2.3: % contribution of sources in temporal domain without (a) and with UMS (b)

| (a) Without UMS | | | |
|-----------------|----------------|-------------|-------|
| source | % contribution | | |
| | HTIMIT | OGI Stories | TIMIT |
| phone | 6.1 | 3.7 | 7.3 |
| speaker | 11.9 | 27.4 | 16.2 |
| channel | 11.5 | | x |
| residual | 70.2 | 68.8 | 76.4 |

| (b) With UMS | | | |
|--------------|----------------|-------------|-------|
| source | % contribution | | |
| | HTIMIT | OGI Stories | TIMIT |
| phone | 7.3 | 4.7 | 7.7 |
| speaker | 9.2 | 7.5 | 10.8 |
| channel | 0.9 | | x |
| residual | 82.6 | 87.6 | 81.3 |

information. We observe that the nature of this information is similar to the information calculated in [30]. This shows that the results of MANOVA are robust to the deviation from normality.

2.8.1 Hierarchical ANOVA

Sun et. al. [15] used hierarchical ANOVA to analyze variability in TIMIT using nine factors: broad phone category, phone unit, phone-in-context, gender, dialect, speaker, token of one speech unit, sub-segment with a token and frame with each sub-segment of the token. The features are 8 Mel frequency cepstral coefficients (MFCCs) after mean and variance normalization. The contribution of the factors is measured using equation 2.4. The results are presented in terms four combined factors: 1) phone - phone broad class + phone unit, 2) context, 3) speaker - gender, dialect, speaker, 4) token - token of one speech unit + sub-segment with a token + frame with each sub-segment of the token. Results conclude that 34.1 % variation is due to different phones, 11.6 % variation is due to different speakers, 27.9 % variation is due to differences in the phonetic context, and remaining 26.5 % variation is among the sub-segments within each segment.

These factors map to our three factors as follows. Channel does not map to anything as TIMIT was recorded through single channel. The phone factor maps to the phone

factor from our study. Speaker factor maps to phone-specific speaker factor from our study. Combination of phonetic context and variation among the segments contribute to the residual variability. We perform MANOVA using similar features on TIMIT database and results show that the phone variability is 31 %, speaker variability is 12 % and residual variability is 57 %. Thus the results from the two analysis are similar.

2.8.2 Mutual Information

Yang et. al. [30] proposed an information-theoretic approach for analyzing variability in speech. They showed that sources in speech do not have a Gaussian distribution. They also noted that if variability is measured as mutual information then it is invariant to any (linear or nonlinear) transformations of the feature space.

The following sub-sections give a brief description of this approach. We describe the terms used in this approach and a kernel-based method used to model the feature distribution. The results from MANOVA are converted to mutual information and are compared with the results obtained in [30].

Information, mutual information and joint mutual information

If $X = \{x\}$ is a random variable with the probability distribution $p(x)$ then the information in x is given by entropy,

$$H(X) = - \sum_{x \in X} p(x) \log p(x).$$

Entropy is measured in bits. Let $Y = \{y\}$ denote another random variable with the probability distribution $p(y)$. Mutual information between X and Y is a measure of reduction in the uncertainty of X after observing Y . It is defined as

$$I(Y; X) = H(X) + H(Y) - H(X, Y) = H(Y) - H(Y/X) = H(X) - H(X/Y) \quad (2.5)$$

If $Z = \{z\}$ is another random variable then the joint mutual information (JMI) between Y, Z and X is the reduction in the uncertainty of X after observing both Y and Z . It is defined as

$$I(Y, Z; X) = H(Y, Z) - H(Y, Z/X) \quad (2.6)$$

In the context of speech recognition, we refer to X as a set of classes. $H(X)$ is based on the prior probability of these classes. Y and Z represent two measurements made from the speech signal. Subsequently, $I(Y; X)$ and $I(Y, Z; X)$ represent the improvement in the separability of the classes after modeling classes using Y and Z .

MI and JMI using kernel-based methods

In this method [7], the distribution of y is divided into K bins. The probability of samples in each bin is approximated by $p_i = r_i/N$ where r_i is the number of times $y = i$ in N samples. The entropy is estimated as

$$H(Y) = - \sum_{i=1}^K p_i \log p_i.$$

The conditional entropy and the mutual information is calculated similarly. Note that this method does not make any assumption about the distribution of y . The drawback is that maximum value of $H(Y)$ using this method is $\log(N)$ whereas the true value of $H(Y)$ can only go up to $\log(K)$. Different techniques have been proposed to correct the estimates of the entropy.

MI and JMI using MANOVA

Computation of MI can be simplified by assuming a parametric distribution for Y . In our case, Y is assumed to have a Gaussian distribution with mean μ and variance σ^2 , i.e., $p(y) \sim N(\mu, \sigma^2)$. So the differential entropy can be written as

$$h(Y) = \frac{1}{2} \log(2\pi e \sigma^2).$$

Similarly, if the distribution of Y within each class is assumed to have a Gaussian distribution with variance σ_i^2 then $h(Y/x = i) = \frac{1}{2} \log(2\pi e \sigma_i^2)$. Therefore, the mutual information can be written as,

$$\begin{aligned} I(X; Y) &= \frac{1}{2} \log(2\pi e \sigma^2) - \sum_{i \in X} p(i) \frac{1}{2} \log(2\pi e \sigma_i^2) \\ &= \frac{1}{2} \{ \log(\sigma^2) - \sum_{i \in X} p(i) \log(\sigma_i^2) \}. \end{aligned} \quad (2.7)$$

Note that this formulation is accurate when the total and the within-class distributions of features are Gaussian distributions. In general, the results are accurate up to the second-order statistics for non-Gaussian distributions.

This expression is similar to an expression obtained using analysis of variance (ANOVA) as follows. The total variance is decomposed using ANOVA as,

$$\sigma^2 = \sigma_{ac}^2 + \sigma_{wc}^2$$

where σ_{ac}^2 is the across-class variance and $\sigma_{wc}^2 = \sum_{i \in Y} p(i) \sigma_i^2$, is the within-class variance. Therefore

$$\sigma_{ac}^2 = \sigma^2 - \sum_{i \in Y} p(i) \sigma_i^2 \quad (2.8)$$

Comparing equations 2.5, 2.7 and 2.8, we can see that mutual information, $I(X; Y)$, can be compared with σ_{ac}^2 ; and conditional entropy, $h(Y/X)$, can be compared to σ_{wc}^2 . Conditional entropy is the logarithm of geometric mean of the individual class variances and within-class variance is the arithmetic mean of the individual class variances.

Similar derivation can be made for computing joint mutual information. Let Σ be the covariance matrix of $\{Y, Z\}$ and Σ_i be the within-class covariance. The expression for JMI and the corresponding equation for MANOVA becomes

$$I(Y, Z; X) = \frac{1}{2} \log(|\Sigma|) - \sum_{i \in X} p(i) \frac{1}{2} \log(|\Sigma_i|)$$

and

$$\Sigma_{ac} = \Sigma - \sum_{i \in Y} p(i) \Sigma_i$$

This shows that information, mutual information and joint mutual information can be computed using the variance/covariance terms derived from MANOVA.

Measurements in Time and Frequency

To compare results, we have used the same database, features and classes used in [30]. Note that absolute MI values cannot be compared across these methods. It has been shown that entropy of a non-Gaussian distribution is overestimated using Gaussian assumption [58]. The same conclusion cannot be applied to MI because MI has two entropy terms. If

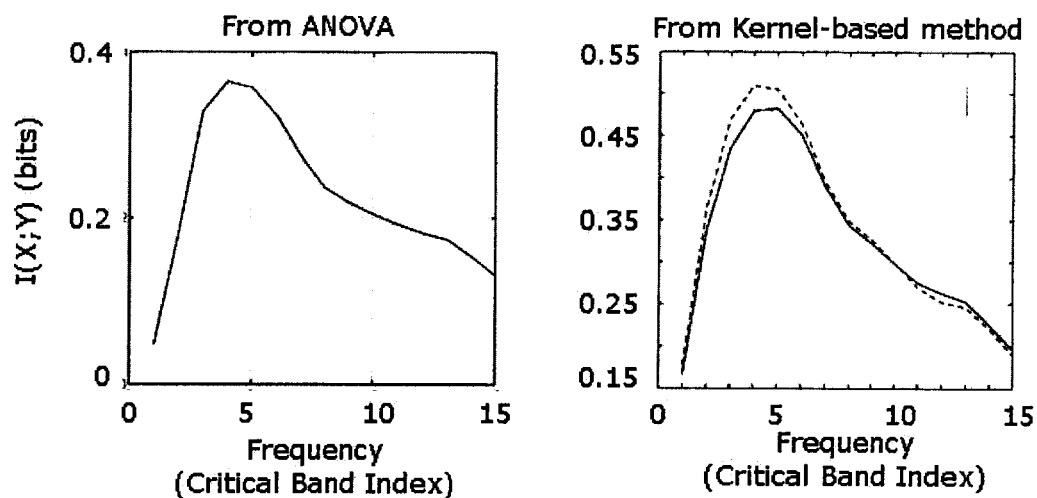


Figure 2.6: Mutual information between frequency band and the phone labels computed assuming Gaussian distribution

the distribution is not Gaussian then the difference in the overestimation of the entropies will give estimate of MI that is greater than or less than the actual MI. Therefore, we compare only the nature of MI in time-frequency domain in the following comparisons.

Distribution of information in frequency

Figure 2.6 shows that MI using kernel-based method [30] $I(X; 5) = 0.5$ bits is different from MI using MANOVA-based method, $I(X; 5) = 0.36$ bits. The relative MI across bands is similar using the two methods. Both these methods indicate that fourth or fifth critical band provides the highest information about different phones.

Distribution of information in time

Similar results are obtained using both the methods (Figure 2.7 and [30]), i.e., information about current frame lasts for approximately 200 ms around the frame.

Additional feature at different frequency but the same time

Both methods (Figure 2.8 and [30]) indicate that the optimal second measurement after the first measurement at 5 Barks, is at 9 barks. However, the gain in the MI differs. Using kernel-based method, gain because of second measurement in spectrum is 0.35 bits. Using

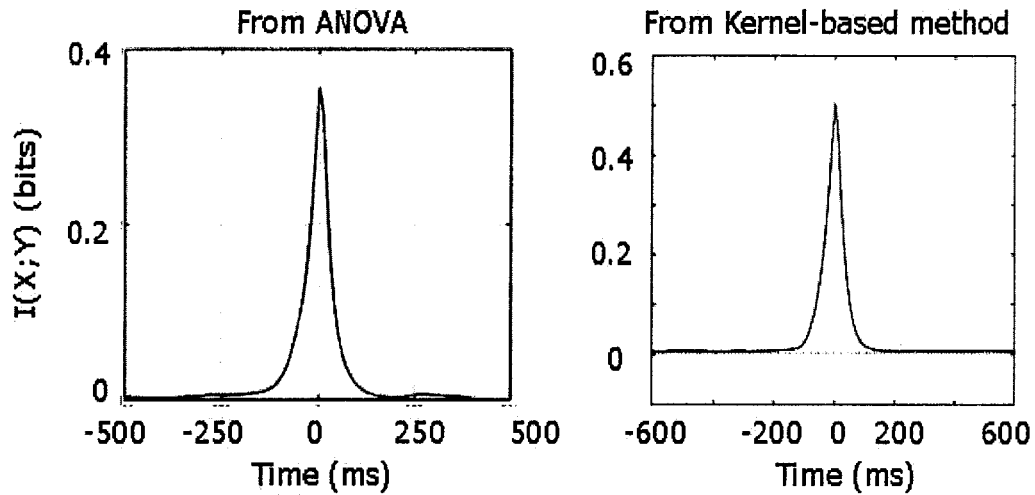


Figure 2.7: Mutual information between feature at time t and the phone label at 5 Barks assuming Gaussian distribution

MANOVA-based method, gain due to second measurement in spectrum is 0.21 bits if the first measurement is fifth critical band and it is 0.26 bits if first measurement is fourth critical band.

Additional feature at different time but the same frequency

1. Kernel-based method ([30]) - The highest gain due to the second feature in time is 0.18 bits. There is asymmetry about the information gain due to second measurement in time - measurement in past is more important than measurement in future. Best measurement in the past is 20 ms before the current measurement and best measurement in the future is 100 ms after the current measurement.
2. MANOVA-based method (Figure 2.9) - Maximum gain due to second feature in the time is 0.07 bits. Second measurement in past gives slightly more information than the measurement in the future. The best second measurement in the past and the future is 50 ms away from the current measurement. No bias is observed the measurement beyond 200 ms on either side does not add significant additional information.

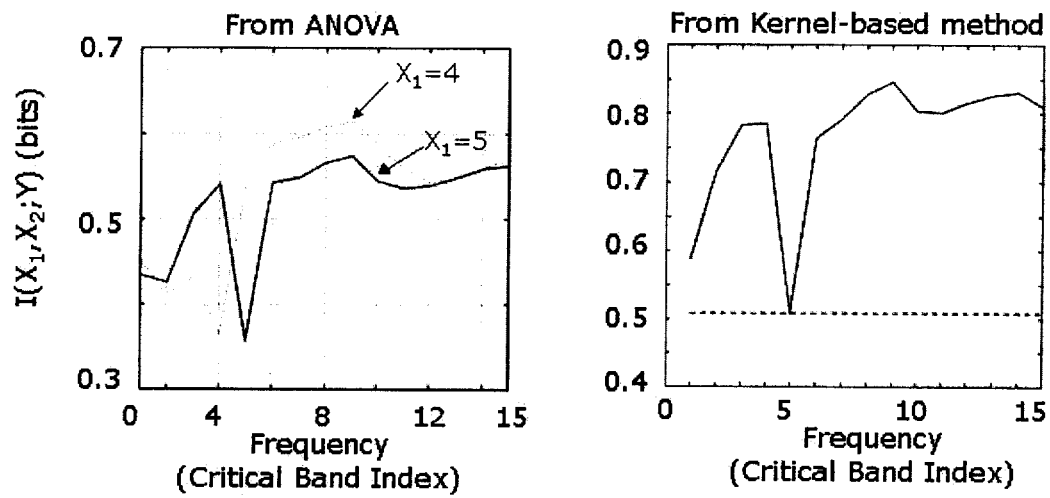


Figure 2.8: Joint mutual information between two features in frequency and the phone labels given that first measurement in frequency is at 4 or 5 Barks assuming Gaussian distribution

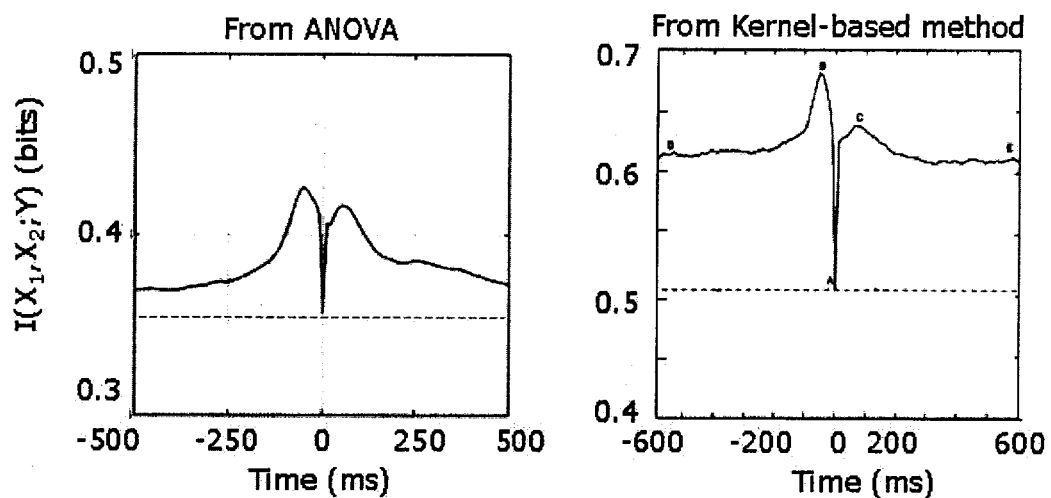


Figure 2.9: Joint mutual information between two measurements in time and the phone labels given that first measurement in time is at $t = 0$. Results are frequency of 5 Barks assuming Gaussian distribution.

Note that a bias is observed using Kernel-based method for the measurements beyond 200 ms. This was due to the fact that the analysis was performed per speaker and per channel. The results, therefore, show not only the mutual information between features and phones but also the mutual information between features and phones, speakers and channels. The higher MI beyond 250 ms in the past and in the future shows the MI between features and speakers and channels. In the MANOVA-based method, we perform the analysis over all the speakers and channels. Our results show minimum bias at 500 ms around the center. Also, note that our results are similar to the speech recognition experiments performed using the fifth band than those obtained using kernel-based method [30].

2.9 Conclusions

This chapter described multivariate analysis of variance (MANOVA) as a technique to analyze variability in speech. It was assumed that variability in speech can be divided into four types - 1) phone variability, 2) speaker variability, 3) channel variability, and 4) residual variability. The sources of first three variabilities were named as phone, speaker and channel respectively, and the residual variability was attributed to the unaccounted sources. Variability in the most commonly used features was decomposed in spectral and temporal domains. Different formulations were proposed to quantify the variability in each domain because spectral domain is bounded by sampling frequency and temporal domain is not bounded. Results showed that phone variability is the highest variability in spectral domain and it spreads for 250 ms around the current time frame in temporal domain. Results also showed significant interactions between phones and speakers. Finally, these were observed to be consistent across different databases and different data-sets.

Results from this study were compared with two other studies. First study [15] analyzed variability in TIMIT using nine factors. We mapped these factors to the three factors from our study, and obtained comparable results. Second study [30] estimated the two most significant measurements in time-frequency domain. It used mutual information to characterize the variability in speech because the distribution of features was observed to be non-Gaussian. We showed that results of MANOVA can also be interpreted as

mutual information and obtained similar results for the most significant measurements in time and frequency domains. Thus MANOVA is robust to deviations from normality [20, 29, 10] and it can be also considered analysis of information.

The main difference between [15, 30] and this study is that we use the analysis to compare different feature-sets. We show that variabilities feature-sets is related to their speech and speaker recognition performances. Further, we also relate MANOVA to LDA where it is shown that results of MANOVA can be used to obtain a robust set of features for speech and speaker recognition tasks. Some of the minor advantages of this method over the previous methods are: 1) this method is computationally less expensive, and 2) this method also measures the interaction between different factors.

Note that analysis of variance in speech is analysis of features extracted from the speech signal. The percentage of variability due to each source is a measure of the sensitivity of the features to different types of variations. In the subsequent chapters, we will show that the performance of the features in speech and speaker recognition task depends on the sensitivity of the features to these sources. Therefore results of MANOVA can be used to study the suitability of the features for a give task.

Chapter 3

MANOVA as a Diagnostic Tool

In the previous chapter, we decomposed total variability in the features into four types: variability due to phones, variability due to speakers, variability due to communication channels, and the residual variability due to the unaccounted factors. Note that these types are chosen because we intend to study the features for speech and speaker recognition tasks. In this chapter, we show that the relative contribution of the variabilities in the features correlates with their performance on the speech and speaker recognition tasks.

The chapter is organized as follows. In Section 3.1, we measure the speaker variability in the features for broad phonetic categories. We show that the speaker variability is related to the speaker recognition performance using these categories. In Section 3.2, we use these results and show improvements in the speaker recognition system on NIST speaker recognition task. Section 3.3 compares relative phone variability with speech recognition performance across different feature-sets and shows that they are related. Section 3.4 gives summary and conclusions of this work.

3.1 Speaker Variability and Speaker Recognition Performance

In chapter 2, we showed that there is more interaction between phones and speakers than between phones and channels. This means that speaker characteristics have a larger variation across different phones than channel characteristics [53]. In this section, we compute relative speaker variability in the features for different phone categories, without explicit speaker recognition experiments. We show that the speaker variability is related

to to the speaker recognition performance obtained using these categories.

Several researchers have pursued a phone-based approach to speaker verification/recognition [3, 50, 1, 36]. Eatock and Mason [3] have shown that nasals and vowels provide best speaker recognition performance, followed by fricatives, affricates, and approximants, with stops providing the worst performance. Gupta and Savic [50] used four broad phoneme categories - voiced, fricative, nasal, and plosive - for the speaker recognition. They concluded that for the speaker recognition task, plosive are least effective whereas the voiced phonemes and fricatives are most effective broad classes. Paris and Carey [1] showed that the log likelihood ratio of different phonemes can be used to characterize their speaker discriminating ability. They concluded that a subset of phones - front vowels, voiced fricatives and nasal - outperforms the complete set of phones on the text-independent speaker recognition task. Rodriguez-Lunares and Garcia-Mateo [36] used three broad phoneme classes - voiced, unvoiced and transitions - for speaker recognition. Their results showed that the class of voiced phonemes is the most effective class for speaker recognition.

The main difference between the previous research and this work is that we are investigating into the nature of speaker variability without running explicit speaker recognition experiments. We are using MANOVA to estimate the importance of broad categories and verifying the results using speaker recognition experiments.

3.1.1 Feature-set

From HTIMIT database, we use data from 4 electret and 4 carbon-button handsets in the following experiments. Eighteen MFCCs (C1-C18) are computed from logarithmic energies from 19 filters in 330-3300 Hz frequency range. The energies are computed using 32 ms speech segment and the adjacent segments are overlapped by 22 ms. The features are appended by delta and double-delta features computed over 4 and 9 frames respectively. Each feature element is normalized using the global mean and variance computed over the feature set used for MANOVA.

3.1.2 MANOVA

A set of 1363 files from 137 speakers (93 females and 44 males) that are common across different handsets is used for MANOVA. Table 3.1 shows the phones used in these experiments. MANOVA is performed the same way as described in the previous chapter. Speaker and channel variability within a broad category is calculated as the average speaker and channel variability within the phones in that category, that is,

$$\begin{aligned} \text{speaker variability, } \Sigma_s &= \frac{1}{N^{bc}} \sum_{c \in bc} \Sigma_s^i * N^i \\ \text{channel variability, } \Sigma_c &= \frac{1}{N^{bc}} \sum_{c \in bc} \Sigma_c^i * N^i, \end{aligned}$$

where, Σ_s^i and Σ_c^i are the phone-specific speaker and channel covariances, N^i is the number of frames in each category. The relative speaker variability is computed as $trace(\Sigma_s)/trace(\Sigma_c)$.

Figure 3.1 shows the results of MANOVA for individual phone categories. It shows that vowels, diphthongs, glides and nasals are more sensitive to speaker changes than the other categories. Features for silence and stops show the lowest sensitivity to speaker changes. Similar results are obtained by clustering phones into broad phonetic categories (Table 3.2). They show that features from vowels and diphthongs are the most sensitive categories to speaker changes. They are followed by features from glides and nasals, and fricatives. Features from silence and stops category show lowest sensitivity to speaker changes.

Table 3.1: Phones and their mapping to broad phonetic categories

| Broad category | Phones |
|----------------------------|-----------------------------------------------------------|
| vowels+diphthongs (v+d) | ix ux ih iy eh ae uh uw aa ax axr ah er ao aw ay ow oy ey |
| glides+nasals (g+n) | w y r l m n em en ng nx eng |
| fricatives (f) | f th s h hv sh v dh z dx zh ch jh |
| silence+stops (s+s) | sil epi pcl tcl kcl qcl bcl dcl gcl b d g p t k |

Table 3.2: Relative speaker variability for features of different broad categories

| Category | v+d | g+n | f | s+s |
|------------------------------|------|-----|-----|-----|
| Relative speaker variability | 10.1 | 9.7 | 6.2 | 3.4 |

3.1.3 Speaker Recognition Experiments

The experiments are performed in a universal background model (UBM) and Gaussian mixture model (GMM) framework [14]. Figure 3.2 shows this framework. There are two types of models: universal background model and speaker model. During training part, UBM is trained over many different speakers and different conditions. GMM have been the most popular modeling technique for UBM [14]. Speaker models are typically adapted from UBM using maximum a-posteriori (MAP) adaptation. During testing part, a hypothesized speaker identity is supplied with the input utterance. The log-likelihood of these features is calculated using UBM and hypothesized speaker model. The difference in the likelihoods is compared to a pre-computed threshold (λ). If the likelihood of the speaker model is higher than the likelihood of UBM then the hypothesized speaker is accepted as a target speaker.

The basic setup of the experiment is 1) train a set of speakers on one handset and 2) test them on all 8 handsets, which is repeated for all the 8 handsets. For each experiment, 32 component UBM is trained using features from approximately 42 male and 42 female speakers. The features are first clustered into 32 bins using vector quantization. The means and variances of the bins are adjusted using expectation-maximization procedure. A different set of 88 speakers (44 females and 44 males) that are common across all the handsets is used as the target speakers. Their models are generated by adapting only the means of UBM using the adaptation factor as 16.

HTIMIT has 10 utterances for each speaker where each utterance is approximately 3 secs. Target models are trained using 8 utterances and 2 utterances are used for testing. During testing, the utterances are tested against all 88 speakers – 1 target speaker and 87 impostor speakers. We use 4 carbon-button and 3 electret handsets for the speaker recognition experiments. The el3 handset is excluded from the testing because it does not

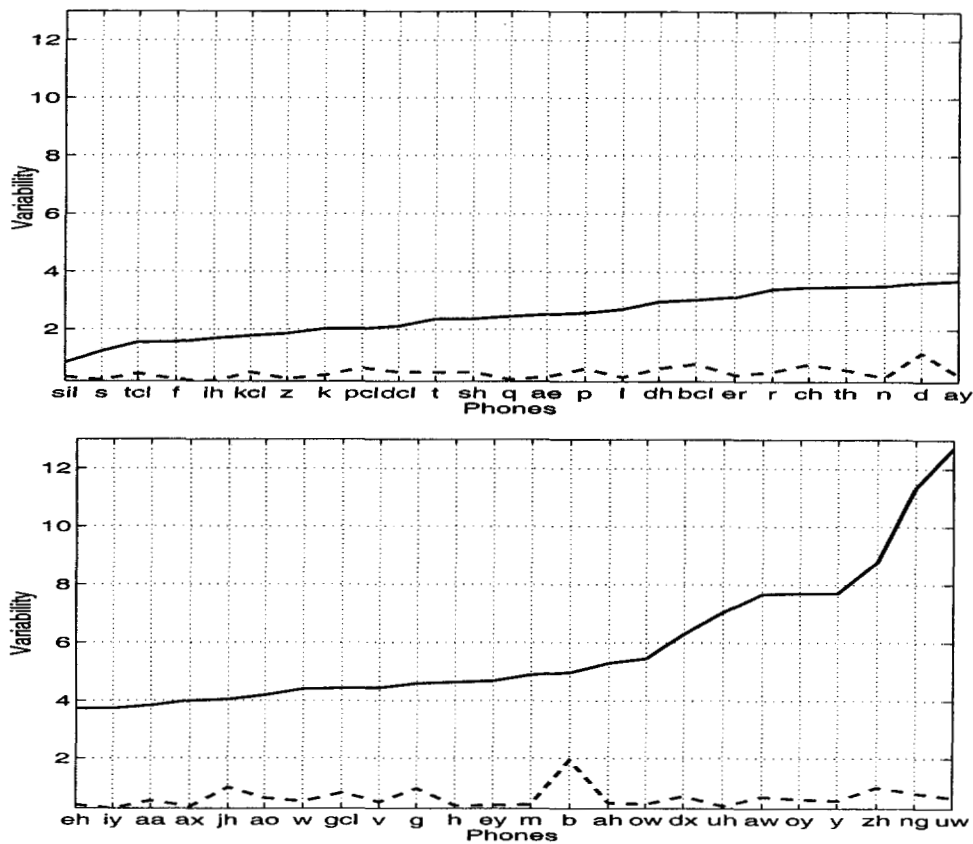


Figure 3.1: Speaker and channel variability within each phone.

have enough data for male speakers.

Table 3.3 speaker recognition results using one category. The results show a similar trend as the results of MANOVA. Vowels and diphthongs give the best results among all categories. This is followed by glides and nasals, and fricatives. Category of silence and stops performs the worst. Note that this trend is same for both matched-handset testing and mismatched handset testing. In these experiments, speaker variability is computed using 1-component GMM and speaker recognition experiments were performed using 32-component GMM. But the speaker recognition results are consistent with the relative speaker variability in the features.

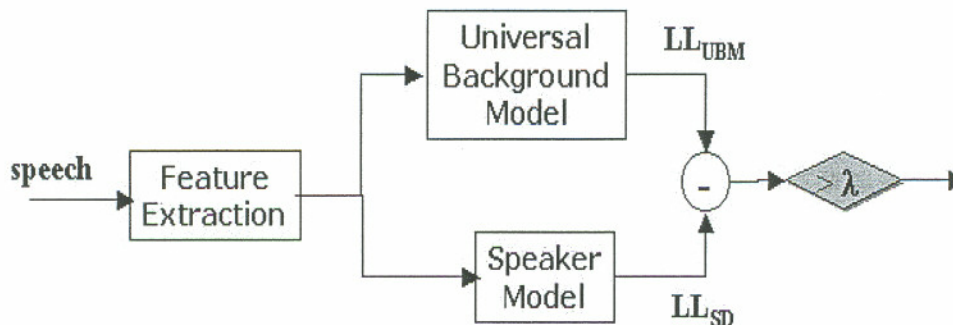


Figure 3.2: Universal background model - Gaussian mixture model (UBM-GMM) framework

Table 3.3: % equal error rates (EERs) using different broad phonetic categories. Same handset condition refers to trials where same handset was used for training and testing. Different handset condition refers to trials where different handset was used for training and testing.

| Broad category | %EER | |
|----------------|--------------|-------------------|
| | Same Handset | Different Handset |
| v+d | 11.6 | 18.3 |
| g+n | 16.1 | 18.9 |
| f | 18.5 | 22.8 |
| s+s | 18.5 | 26.4 |

3.2 Speaker Recognition using Broad Phonetic Categories

Using the results from the previous section, we propose speaker recognition system using broad phonetic categories. It is evaluated on 2001 NIST speaker recognition evaluation task. The broad phonetic category recognition system built using CTIMIT [4]. We present results on this task using different broad categories and show that they agree with the results on HTIMIT database. Further, we show that the performance of the system can be improved using only two broad phone categories - vowels+diphthongs and glides+nasals.

3.2.1 Speaker Recognition Task

This is one-speaker detection task where the goal is to determine if the specified speaker is present in the given speech segment. The evaluation data is drawn from Switchboard-II phase-4 database. There are 174 target speakers - 100 females and 74 males. Each

speaker has approximately two minutes of training data. The task is evaluated using 22418 speaker trials – 2038 trials with target speaker and 20380 trials with impostor speaker. The duration of each trial segment varies between 0 sec and 60 sec. The trials include both “same number” and “different number” trials which refer to same handset and same handset type trials. The results are combined over both the conditions.

The development data is also provided by NIST. It was also drawn from Switchboard-II phase-4 database and does not overlap with the evaluation data. The development data contains 22 female and 39 male speakers with two minutes of training data per speaker. The test data contains 78 segments - 34 female and 44 male.

3.2.2 Recognition System

The system is implemented in UBM-GMM framework as described earlier. A 256 component UBM is trained using the development data. There is one UBM for each gender. The target models are derived with MAP adaptation of UBM where only means are adapted using adaptation factor equal to 16. The features are 15 MFCCs (C1-C15) appended with 15 delta and 15 double-delta coefficients. The features are normalized with the mean computed per utterance and they are processed using a medium-term gaussianizer [32]. This is a state-of-the art speaker recognition system that has consistently performed comparable to the best systems submitted in the evaluation.

The speech recognition system used for broad phonetic transcription is trained using 272 speakers - 136 male and 136 female - from CTIMIT database. Features are 13 MFCCs (C1-C13) appended with 13 delta and 13 double-delta coefficients. They are normalized with the mean computed over the utterance and variance computed over the development database. Each broad phone category is modeled using 3-state, 32-component HMM. During transcription, a simple grammar is used where any category can follow any other category.

3.2.3 Results

Table 3.4 shows the speaker recognition performance obtained using features from one broad phonetic category. The results again show that vowels and diphthongs perform the

best and silence and stops perform the worst among the categories. This is consistent with the results of MANOVA. Note that this result is significant because: 1) unlike HTIMIT, broad phone category transcriptions are generated using ASR system, 2) each category was modeled using 256-component GMM.

Table 3.4: Speaker recognition performance using broad phonetic categories on NIST 2001 speaker recognition task

| Category | v+d | g+n | f | s+s |
|----------|-----|------|------|------|
| %EER | 8.9 | 13.1 | 18.3 | 23.6 |

These results are used to select a smallest set of categories that give the best performance on the speaker recognition task. Table 3.5 shows that the combination of vowels, diphthongs, glides and nasals gives the best performance (Figure 3.3). Note that it is also better than using speech-silence segmentation where speech-silence segmentation is performed using artificial neural network (ANN). ANN was trained using same data that was used to train the transcription system. It used 13 PLP [24] cepstral coefficients appended with 13 delta and 13 double delta coefficients. ANN had 9 frames as input (39×9 input nodes), 500 hidden nodes and 2 output nodes. This gave 8.5 % EER which is similar to the system that ignored the silence and stops category.

Table 3.5: Speaker recognition performance using combination of broad phonetic categories on NIST 2001 speaker recognition task

| Categories | %EER |
|-------------------------|------------|
| v+d,g+n,f,s+s | 9.8 |
| v+d,g+n,f | 8.9 |
| speech-sil segmentation | 8.5 |
| v+d,g+n | 7.8 |
| v+d | 8.9 |

Note that the results of MANOVA show that all the categories have non-zero relative speaker variability. Therefore, the performance of the systems should not degrade when features for some categories are added to the system. In the worst scenario, the performance of the system should remain unchanged if the new data does not contribute to

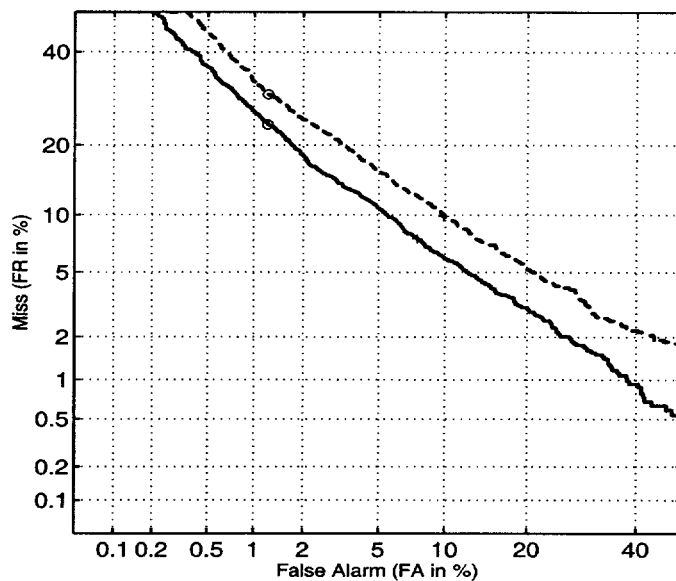


Figure 3.3: Performance of speaker recognition system using all categories (dotted-line) and two broad phone categories (straight-line) on NIST 2001 speaker verification task

speaker recognition. The performance of degradation of the system by adding certain categories is related to modeling efficiency, that is, ratio of Gaussian components to the useful categories. With fewer and most useful categories, these categories are modeled using all the Gaussian components. By adding categories like silence+stops, many components are used to model this category and the ratio of Gaussian components to the useful categories reduces. Note that this result is valid only with the constraint that the size of GMM is constant while adding more data. If the number of components are increased then the drop in the performance will not be significant.

3.3 Phone Variability and Speech Recognition Performance

This section compares the relative phone variability with the speech recognition performance over different feature-sets. The relative phone variability is calculated using MANOVA on HTIMIT database. The experimental setup is same as the one described in 2. The relative phone variability is defined as

$$trace((\Sigma_{total} - \Sigma_{phone})^{-1}\Sigma_{phone}).$$

This is also called as F-ratio obtained using phones as classes. The speech recognition experiments are performed on digit recognition tasks using OGI Numbers and noisy TIDIGIT databases.

3.3.1 Effect of Delta and Double-delta Features

In these experiments, we use three feature-sets with 24 dimensional features. The base features are 24 logarithmic filterbank energies projected on 24 discrete cosine transform (DCT) bases. They are compared to two features-sets- 1) 12 DCT coefficients appended with 12 delta features computed using a window of 4 features, and 2) 8 DCT coefficients appended with 8 delta features using a window of 4 features, and 8 double-delta coefficients computed using a window of 9 frames.

The recognition experiments are performed using OGI Numbers database. It is a continuous digit recognition task. The digits are modeled as sequence of monophones and each monophone is modeled using 5-state, 3-component HMM. A simple grammar is used where any digit can follow any other digit. The performance is measured as percentage word-error-rate (%WER).

Table 3.6 shows the relative phone variability and speech recognition performance. Results show that speech recognition performance improves when the relative phone variability improves.

Table 3.6: Phone variability and speech recognition performance using delta and double-delta features

| Features | Relative phone variability | %WER |
|-----------------------------------------------|----------------------------|------|
| 24 DCT coefficients | 3.9 | 11.9 |
| 12 DCT + 12 delta coefficients | 4.1 | 8.0 |
| 8 DCT + 8 delta + 8 double-delta coefficients | 4.3 | 6.9 |

3.3.2 Effect of Feature Transformations

This section analyzes different processing blocks used in our submission for the AURORA task. Figure 3.4 shows the block diagram of the system. We analyze features at different stages (marked by roman numerals) and measure relative phone variability in the features. This is related to the performance of those features obtained on the AURORA task.

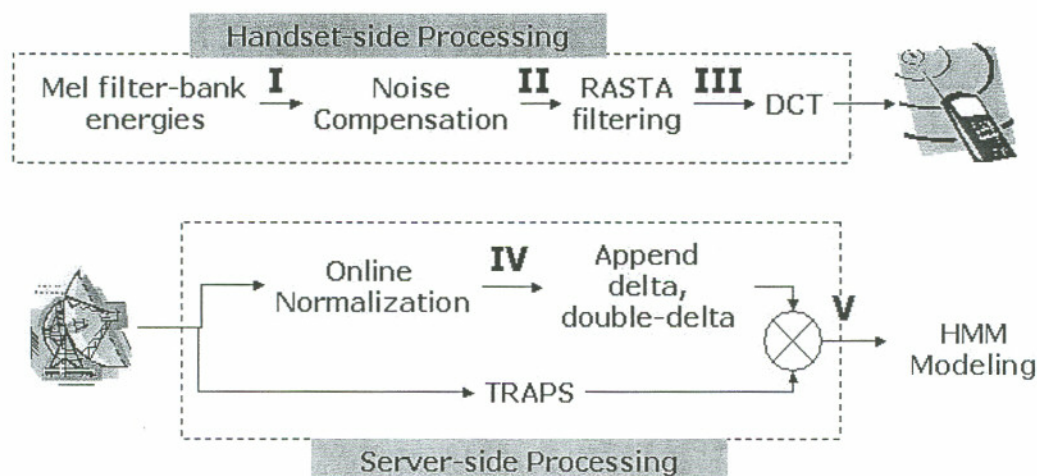


Figure 3.4: Modified QUALCOMM-OGI-ICSI front-end used for AURORA task

A brief description of the task is presented here. Please refer to [5] for a detailed discussion. The task is distributed speech recognition in different languages under different additive noise and convolutive noise conditions. Speech recognition refers to connected digit recognition. In this work, we consider only English digit recognition task. The task was designed using TIDIGIT database where different noises were artificially added to it. The recognizer was fixed by the standards committee where each digit was modeled using whole-word model. The word models were 16-state, 3-component HMMs. The participants were allowed to only change the features under the constraints of minimum bit-rate and minimum latency of 250 ms.

Different feature-sets are extracted using the system as follows:

1. Logarithmic energies from 23 filters on MEL frequency axis (I). They are projected on 15 DCT bases and are appended with 15 delta and 15 double-delta coefficients;

2. Logarithmic energies after noise compensation (II). They are projected on 15 DCT bases and are appended with 15 delta and 15 double-delta coefficients;
3. Cleaned logarithmic energies after RASTA filtering (III). They are projected on 15 DCT bases and are appended with 15 delta and 15 double-delta coefficients;
4. RASTA filtered energies after online mean and variance normalization (IV). They are projected on 15 DCT bases and are appended with 15 delta and 15 double-delta coefficients; and
5. Normalized 45 features appended with 6 TRAPS features and reduced to 45 features using whitening transformation (V).

Note that all the feature sets have 45 features.

For all the features, the training data consists of clean TIDIGIT database and four different additive noises added to TIDIGIT database. The testing data consists of three parts. Test A has same noises used in training set, Test B has different additive noise, and Test C has two different additive noises with two different convolutive noises. The recognition performance is the average WER improvement over previous AURORA standard in these testing conditions.

Table 3.7 shows the relative phone variability computed using 5 feature-sets and their recognition performance on English part of AURORA task. Results show that improvement in the relative phone variability is related to improvement in performance of features. Note that the small improvement in relative phone variability after noise compensation and online normalization is due to the fact that HTIMIT database is less noisy than the noisy TIDIGIT database used in AURORA.

3.4 Conclusions

In this chapter, we related variabilities in different features to their performance on speech and speaker recognition tasks. We measured speaker and channel variabilities in the features for different broad phonetic categories. It was observed that the category of vowels and diphthongs is the most sensitive category for speaker variations. Thus, it

Table 3.7: Phone variability and recognition performance at different stages of proposed AURORA front-end

| System | Relative phone variability | % improvement |
|--------|----------------------------|---------------|
| 1 | 4.8 | 1.3 |
| 2 | 4.9 | 16.3 |
| 3 | 5.1 | 23.4 |
| 4 | 5.2 | 30.1 |
| 5 | 6.3 | 38.6 |

was hypothesized that it must be the most useful category for speaker recognition. The hypothesis was verified using speaker recognition experiments.

We showed that relative phone variability in the features is related to their performance on speech recognition tasks. This was shown using two experiments. In the first experiment, we compared features with and without dynamic features. It was shown that adding delta and double-delta features improve the relative phone variability in the features. They were also shown to improve the speech recognition performance.

In the second experiment, we analyzed the system submitted for AURORA task. The system had different processing stages where each stage was shown to improve the performance of the features. We computed relative phone variability using features at the output of different processing stages. The relative phone variability was shown to improve after these processing stages. Thus, it was shown that relative phone variability in the features is related to their speech recognition performance.

In conclusion, we showed that results of MANOVA can be used to compare different feature-sets for speech and speaker recognition tasks. In the following chapters, we will show that the factor covariances, which were used to compute the relative contributions, can also be used to compute new features which are suitable for speech recognition task.

Chapter 4

MANOVA and LDA

In previous chapters, it was shown that given a set of features, MANOVA can be used to decompose the total variability in the features into four types - phone (inter-phone), speaker, channel and residual. The percentage contribution of these variabilities was shown to be proportional to the performance of the features on the speech and speaker recognition tasks. In this chapter, we will show that the variabilities computed using MANOVA can be used to derive robust features for a given task using a statistical technique, linear discriminant analysis (LDA).

We begin the chapter with an overview of MANOVA in Section 4.1. Section 4.2 describes LDA in general, and Section 4.3 gives an overview of the previous work in speech recognition using LDA. In Section 4.5, we modify MANOVA assuming three-state phone models. Resulting variabilities are used to derive linear discriminants in Section 4.6. It is shown that the new discriminants improve the performance of continuous digit recognition system over those derived using one-state phone models.

4.1 MANOVA

In this thesis, MANOVA is used to decompose the total variability in the features into variabilities due to phones, speakers and channels, and a residual variability. The variability is measured using covariances. The decomposition is performed as

$$\Sigma_{total} = \Sigma_{phone} + \Sigma_{speaker} + \Sigma_{channel} + \Sigma_{residual}.$$

For a given task, all the sources of variability are not important. For example, for speaker independent speech recognition task, Σ_{phone} is useful variability and ($\Sigma_{total} -$

Σ_{phone}) is the harmful variability. For speaker recognition task, $(\Sigma_{phone} + \Sigma_{speaker})$ is the useful variability and $(\Sigma_{channel} + \Sigma_{residual})$ is the harmful variability. The reason for not using only $\Sigma_{speaker}$ as the useful variability is as follows. $\Sigma_{speaker}$ represents two types of variabilities - global speaker variability and phone-specific speaker variability. Conventionally, the first type is eliminated using UMS, and $\Sigma_{speaker}$ represents only phone-specific speaker interactions. Maximizing phone-specific speaker variability implies maximizing both phone and speaker variabilities.

4.2 Linear Discriminant Analysis

LDA is a statistical technique used for feature selection [33], that is, for finding a subspace in the original feature space that contains maximum useful variability and minimum harmful variability for a given task. If the transformation is a linear combination of the input variables then the analysis is also referred as ‘‘Fisher’s discriminant analysis’’. The analysis is performed as follows. Let $\{X_{ij}, Y_{ij}; 0 \leq i \leq N, 0 \leq j \leq M_i\}$ be M ($=\sum_i M_i$), tuples of feature vectors X_{ij} and class labels Y_{ij} . There are N classes present in the data, and there are M_i vectors per class. the task is find a subspace that maximizes separation between the classes while minimizing the variability within classes. Note that the estimate of separation between classes is a useful variability and the average variability within classes is a harmful variability. The optimization criterion is defined as

$$J(e) = \frac{e^t A_c e}{e^t W_c e},$$

where e is a set of vectors that maximize $J()$, and A_c and W_c is defined as,

$$A_c = \sum_{i=1}^N \frac{M_i}{M} (\bar{X}_i - \bar{X}_{..})^t (\bar{X}_i - \bar{X}_{..})$$

$$W_c = \sum_{i=1}^N \frac{M_i}{M} \left\{ \frac{1}{M_i} \sum_{j=1}^{M_i} (X_{ij} - \bar{X}_i)^t (X_{ij} - \bar{X}_i) \right\}$$

\bar{X}_i is a an estimate of class mean, $\bar{X}_{..}$ is an estimate of global mean, and the equation in $\{ \}$ is an estimate of covariance of each class. The solution to this optimization is similar to the generalized eigenvector equation [33],

$$A_c E = W_c E \Lambda$$

where E is a matrix whose columns are e_i and Λ is a diagonal matrix of eigenvalues. If W_c^{-1} exists then this can be also written as

$$W_c^{-1}A_cE = E\Lambda \quad (4.1)$$

The leading eigenvectors of E define the new feature space that maximizes A_c while minimizing W_c . We refer to these eigenvectors as linear discriminants (LDs). The new set of features are obtained by projecting original features on LDs,

$$\tilde{X} = XE.$$

We refer to the new features as discriminant features.

Note that LDA makes the same assumptions about the data as MANOVA. It assumes that the total distribution of features is Gaussian, and the distribution of the class means is also Gaussian. Further, it assumes that the distribution within-each class is the same, and it is also Gaussian. Thus results of MANOVA can be used as input to LDA to derive discriminant features. In the next section, we will review the previous research that have used LDA in speech recognition.

4.3 LDA in Speech Recognition

Many researchers have used LDA in speech recognition [41, 39, 38, 35, 34, 44, 26, 8, 56]. This work can be divided into three groups - LDA in spectral domain, LDA in temporal domain, and LDA in joint time-frequency domain. The first two groups are described in this chapter. The third group is reviewed in the next chapter.

4.3.1 Spectral Domain

LDA in spectral domain refers to the analysis using short-term spectral energies or its linear transformations (for example, DCTs) as features. Hermansky et. al. [26] used logarithmic energies from 15 Bark filters as features. Hunt et. al. [40] used DCT coefficients from spectral energies as features. The classes used in the analysis have varied from a set of context-independent phones [26] to a set of states of HMM [40]. Apart from improving the performance of ASR system, LDs have also been studied for nature of discriminability

in spectral domain. Malayath [42] has shown that linear discriminants from logarithmic FFT spectrum analyze low frequency part with higher resolution than high frequency part. This is similar to the MEL/BARK frequency scale that has more samples in low frequency part than higher frequency part. This work is an extension of [42]. We refer to LDs in spectral domain as spectral LDs (SLDs).

4.3.2 Temporal Domain

LDA in temporal domain refers to the analysis performed using temporal trajectories of spectral energies at a particular frequency as features. Avendano et. al. [9] and van Vuuren et. al. [56] have derived discriminant base functions from TIMIT, Switchboard, and NTIMIT databases. Avendano et. al. [9] used 10 broad phonetic categories as classes and van Vuuren et. al. [56] used context-independent monophones as classes. The bases are interpreted as filters. The filters were similar across different frequency bands. The frequency response of the first filter was similar to original RASTA [28] filter, and the frequency response of second and third filters were similar to the RASTA filter convolved with a delta and double-delta filters respectively. This work is an extension of [56]. We refer to LDs from temporal domain as temporal LDs (TLDs).

4.4 Effect of Phone Classes on LDA

In this section, we show an interesting effect of phone classes on the resulting TLDs. We show that when two classes, that always follow each other, are used in LDA, the resulting discriminants have a delta-like filter.

A commonly occurring case of two classes that follow each other, is closures and bursts within stops. We use two sets of classes where one set has closures and bursts modeled together as stops and another set has closures and bursts modeled separate as two different classes. Figure 4.1 shows first three TLDs for the two cases. First and second discriminants are similar in these two cases. They are approximately symmetric band pass filters. The third filter, however, is a symmetric filter when closure and burst are modeled together and an antisymmetric filter when they are modeled separately.

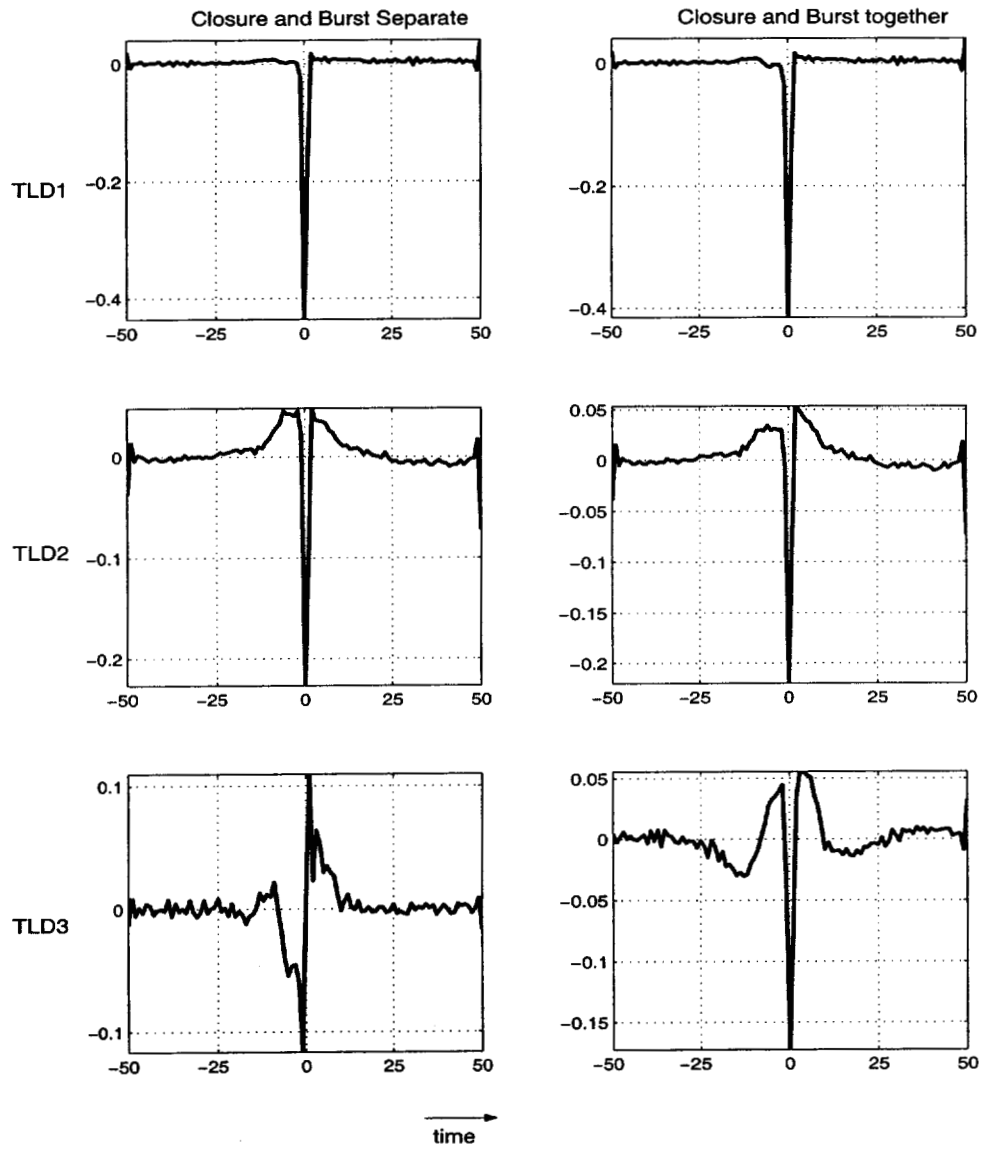


Figure 4.1: Effect of separation of closure and burst on the design of TLDs.

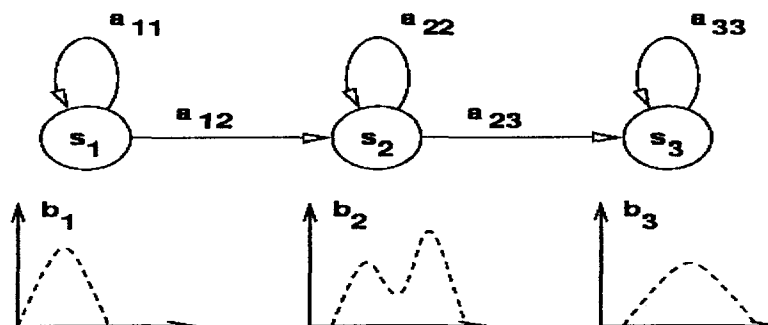


Figure 4.2: Block diagram of HMM where states follow each other in a sequence.

This result is important because delta filter has been used in speech recognition for a more than a decade [48, 49]. It has been proposed based on results of perceptual experiments where the transitions were shown to be important in speech perception. Using these experiments, we show that this filter can be derived from the data when one class follows the other. In general, speech recognition systems model phones as multiple-state HMMs where the states within a phone always follow each other (Figure 4.2). Therefore we expect that discriminants derived using the states as classes will give delta-like filter as one of the discriminants.

4.5 Three-state MANOVA

MANOVA assumes that the distribution of means of different phones is a Gaussian distribution. It also assumes that distribution of features within each phone is a Gaussian distribution, that is, one-state one-component HMM. Linear discriminants derived using results of MANOVA are optimal when phones are modeled as one-state one-component HMMs in recognition experiments. In recognition experiments, however, the phones are typically modeled using multiple-state HMMs, that is, each phone segment is divided into multiple sub-phone segments along time and the segments are modeled independently. To make MANOVA similar to the speech recognition experiment setup, we propose MANOVA using sub-phones instead of phones. In our work, the sub-phones are obtained by dividing the phone segment uniformly into three parts which are analogous to three states of HMM. We also refer to each sub-phone as a *state*. Each state is used as a class for MANOVA.

We refer to this analysis as three-state MANOVA and, with the same analogy, refer to the previous analysis as one-state MANOVA.

The original one-state model for MANOVA was

$$\Sigma_{total} = \Sigma_{phone} + \Sigma_{speaker} + \Sigma_{channel} + \Sigma_{residual}.$$

The new model is

$$\Sigma_{total} = \Sigma_{statephone} + \tilde{\Sigma}_{speaker} + \tilde{\Sigma}_{channel} + \tilde{\Sigma}_{residual},$$

where $\Sigma_{statephone}$ is the variability across different states of phones; and $\tilde{\Sigma}_{speaker}$, $\tilde{\Sigma}_{channel}$ and $\tilde{\Sigma}_{residual}$ are the new speaker, channel and residual variabilities. Note that total variability Σ_{total} is same in both cases.

The variabilities are computed using the same experimental setup described in Chapter 2. The features are preprocessed using UMS. $\Sigma_{statephone}$ is the covariance of the means of states of different phones. Since the results are presented after UMS, the new speaker and channel variabilities represent the interaction between speakers and channels, and different states within a phone.

Tables 4.1 and 4.2 show the contribution of variabilities computer using one- and three-state phone models in spectral and temporal domains. The results are interpreted as follows. $\Sigma_{subphone} - \Sigma_{phone}$ is the average within-state variability in each phone. Results show that within-state variability in temporal domain is comparable to Σ_{phone} . In spectral domain, within-state variability is only a small percentage of Σ_{phone} . $\tilde{\Sigma}_{speaker} - \Sigma_{speaker}$ is the interaction between speakers and different states within a phone, and $\tilde{\Sigma}_{channel} - \Sigma_{channel}$ is the interaction between channels and different states within a phone. Results show that in both domains, the interaction between speakers and states within phones is higher than the interaction between channels and states within phones. $\Sigma_{residual} - \tilde{\Sigma}_{residual}$ is the part of residual variability explained by different states within a phone. The difference in the contribution of residual variability shows that 3.3% variability in spectral domain and 6.4% variability in temporal domain is explained by modeling states within a phone.

Table 4.1: Contribution of variabilities in spectral domain using 1-state and 3-state phone models

| source | % contribution | |
|----------|----------------|-------------|
| | one-state | three-state |
| phone | 58.1 | 59.8 |
| speaker | 6.9 | 9.1 |
| channel | 2.9 | 2.5 |
| residual | 31.9 | 28.6 |

Table 4.2: Contribution of variabilities in temporal domain using 1-state and 3-state phone models

| source | % contribution | |
|----------|----------------|-------------|
| | one-state | three-state |
| phone | 7.3 | 9.9 |
| speaker | 9.2 | 13.0 |
| channel | 0.9 | 0.9 |
| residual | 82.6 | 76.2 |

Figure 4.3 shows Σ_{phone} and $\Sigma_{statephone} - \Sigma_{phone}$ in spectral and temporal domains. It shows that within-state variability in spectral domain is small compared to one-state phone variability. In temporal domain, within-state variability is comparable to 1-state phone variability. It has two peaks around 50 ms from the current state. Assuming that length of a phone is approximately 80 frames, these peaks represent variation at the center of the neighboring phones. This shows the effect of the current phone lasts beyond its boundaries.

4.6 Three-state LDA

Linear discriminants are derived from the results of MANOVA as follows. The across-class and within-class covariances are defined in terms of $\Sigma_{statephone}$ and Σ_{total} as,

$$A_c = \Sigma_{statephone}$$

$$W_c = \Sigma_{total} - \Sigma_{statephone}$$

The linear discriminants are obtained using equation 4.1. Note that SLDs from 3-state phone models are almost the same as SLDs from 1-state phone models, but TLDs are

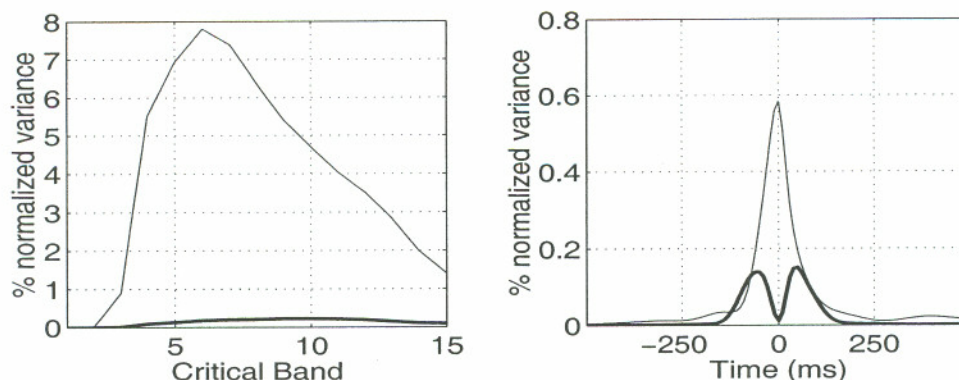


Figure 4.3: Phone variance (thin line) and average within-state variance (thick line) in spectral and temporal domain

different for these two models. Figure 4.4 compares TLDs obtained using phones as classes (case 1) with TLDs obtained using states within phones as classes (case 2). As shown before, first TLD is approximately symmetric band-pass filter, which is similar in both cases. Further, it can be shown that the same discriminant is obtained by using sonorants and obstruants as two classes. Therefore this basis discriminates between sonorants and obstruants.

The second TLD from case 1 is similar to third TLD from case 2. This is also a symmetric band-pass filter. It can be shown that same filter is obtained when vowels, diphthongs, glides, and nasals are used as one class; and schwas as the other class. Therefore, this basis discriminates within sonorants.

The third TLD from case 1 is different from second TLD from case 2. Third TLD from case 1 is a combination of symmetric and anti-symmetric filter, where the symmetric filter is dominant. Second TLD from case 2 is an anti-symmetric filter. The comparison shows that when LDA is performed using phones as classes, third TLD has an antisymmetric component. When LDA is performed using states within phones, the anti-symmetric basis is the second most dominant basis. This shows that the delta-like filter discriminates between states within phones and it becomes dominant when the states are modeled explicitly.

The LDs are evaluated on continuous numbers recognition task. The experiment setup

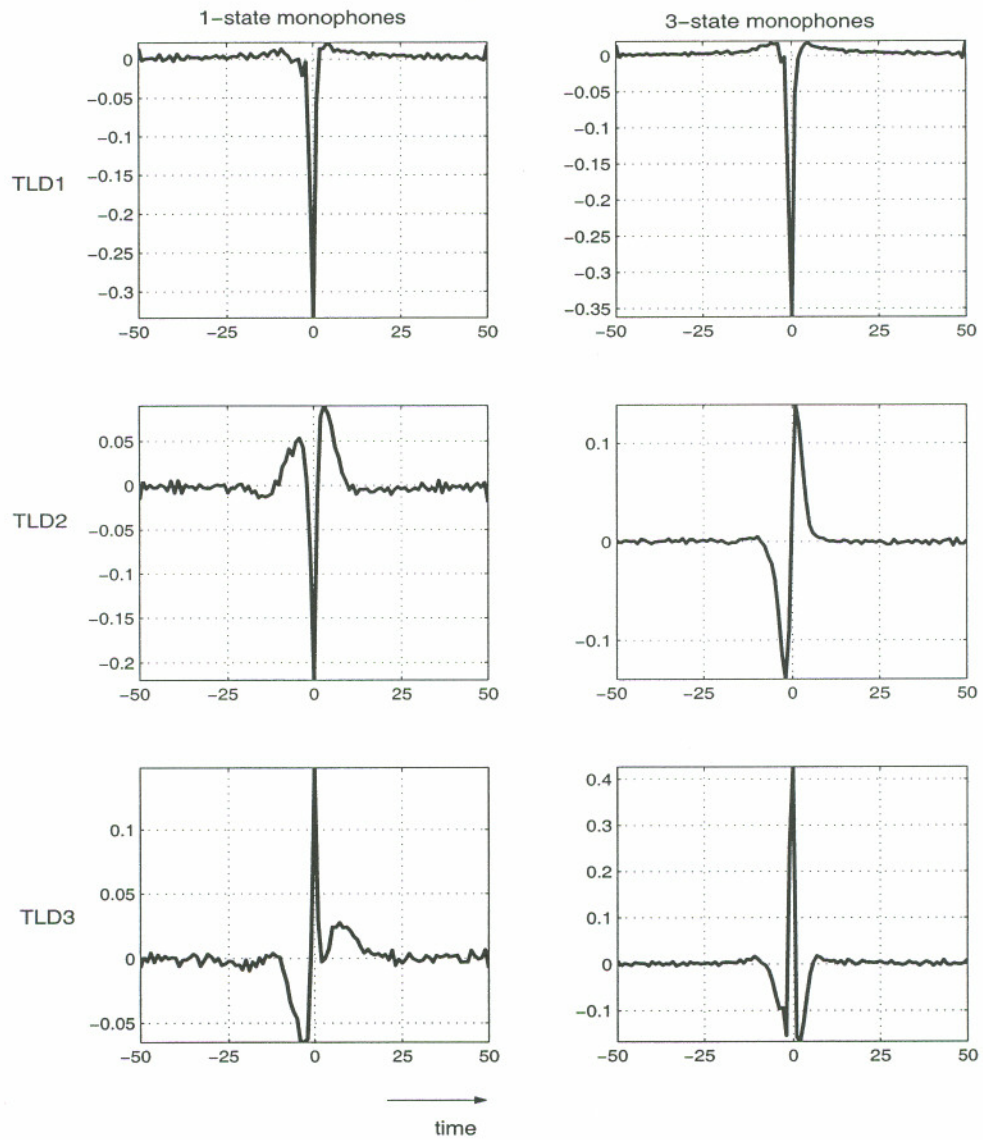


Figure 4.4: Effect of 3-state design on TLDs

Table 4.3: % WER using one-state and three-state spectral discriminants (SLDs) and temporal discriminants (TLDs)

| System | Features | %WER |
|--------|---------------------------------------------|------|
| 1 | one-state SLDs+ Δ + $\Delta\Delta$ | 6.7 |
| 2 | three-state SLDs+ Δ + $\Delta\Delta$ | 6.6 |
| 3 | DCT + one-state TLDs | 6.3 |
| 4 | DCT + three-state TLDs | 5.7 |

is same as described in chapter 3. Table 4.3 shows the word error rates (WERs) using four systems, each using 24 features. Systems 1 and 2 use spectral discriminants appended with 8 delta and 8 double-delta coefficients computed using 4 and 9 frames. Systems 3 and 4 use 8 DCT coefficients filtered using temporal discriminants. Results show that three-state SLDs perform similar to one-state SLDs. The three-state TLDs, however, perform significantly better than one-state TLDs.

4.7 Conclusions

In this chapter, we showed that results of MANOVA can be used for feature extraction using LDA. It was noted that both MANOVA and LDA make same assumption about the data, and results of MANOVA can be used as input to LDA. For speech recognition task, we used the estimate of phone variability as across-class covariance and the estimate of remaining variability as the within-class covariance.

It was observed that when the classes follow one another, one of the resulting LDs is a delta-like filter. This was shown by estimating LDs from two cases: 1) when burst and closure within a stop is modeled separately; 2) when burst and closure within a stop is modeled together. First case showed third TLD as an anti-symmetric – delta-like – filter. Second case shows that all the three TLDs are approximately symmetric filters.

Note that the case, where classes follow one another, is related to HMMs used in speech recognition. In typical speech recognition systems, phones or sub-phone units are modeled as multiple state HMMs. The states in HMMs are in a sequence so they also follow one another. To make our analysis similar to the recognition setup, we modified MANOVA and LDA by dividing phones uniformly into three parts and using each part (referred to

as a **state**) as a class.

Results of MANOVA showed that contribution of within-state variability in spectral domain is lower than contribution of phone variability. In temporal domain, contribution of within-state variability is similar to the contribution of phone variability. Results of LDA showed that, using states within phones as classes, the second most important LD is an anti-symmetric filter. Thus we conclude that this filter, which is similar to the delta filter, discriminates between different states within phones.

LDs, using these state-phone models, were evaluated on the continuous numbers recognition task. SLDs did not show any improvement, but TLDs showed significant improvement in the performance over one-state design. We have also used similar procedure for deriving TLDs on AURORA phase-I task. The task and the system are described in [55]. We observed that one-state TLDs gave 15% improvement over baseline and one-state TLDs gave 30% improvement over the baseline features.

In this chapter, we discussed LDA in spectral and temporal domain. In the next chapter, we will perform LDA in joint time-frequency domain. LDs in joint domain are also obtained using three-state phone models because discriminants from three-state phone models have consistently performed better than the discriminants derived using one-state phone models.

Chapter 5

Two-Dimensional Discriminants for ASR

We showed in Chapter 2 that information about current phone lasts beyond its boundaries, into the neighboring phones. The goal of this work is to incorporate a longer time span into the design of features for speech recognition. A trivial way of using the longer time span is to use a wider block of spectrogram - 101x15 points - as features. However, there are a few problems with these features. First, these features are highly correlated. This can be seen by computing condition number - ratio of largest eigen value to the smallest eigen value. The condition number using OGI Stories database - 500,000 frames - is of the order of 10^7 . Approximately 50 principal components capture 90% of the total energy, and approximately 350 principal components capture 99% of the total energy. This clearly shows that the features are not independent, and most of the information in the feature resides in a lower dimensional subspace. Second, if we assume that all the dimensions are important for recognition then we will have to estimate 1515 dimensional covariance matrices to model the distribution of features. By curse of dimensionality [47], we will need very large number of samples - at least $O(2^n)$ ¹ - to estimate them which is not feasible. We need to estimate a smaller set of features from this block that captures most of the useful information for speech recognition. In this work, the reduced set of features are obtained using LDA in joint time-frequency domain. The resulting linear discriminants are called two-dimensional² linear discriminants (2DLDs).

This work extends the previous work [34, 44, 59, 35] on 2DLDs where the discriminants

¹ Assuming that minimum of two samples are needed for estimating mean and variance of each dimension

² "Two-dimensional" refers to spectral and temporal dimensions.

are derived using a smaller temporal context of around 9 frames. In [34], a block of 9 frames of 20 dimensional spectral features was used as input to LDA. Discriminant features after projecting these features on 50 LDs were used in the recognition experiments. In [35], the process of deriving LDs was divided into two parts. First, LDs were derived using 24 cepstral coefficients. Spectral discriminant features were obtained by projecting 24 cepstral coefficients on 24 LDs. In second part, temporal discriminants were derived independently for each cepstral component. The 60 most LDs were chosen in temporal domain and spectral discriminant features were filtered using them. It was reported that this method showed improvements over the previously proposed method [34]. In [44], multi-resolution features in time-frequency domain were proposed. The temporal span of the original features was limited to block of 9-17 frames. In [59], a block of 3×3 elements from time-frequency domain was used as in input to LDA. This work is an extension of [34, 35], where we show that the performance of the recognition system improves by incorporating longer time-span.

Section 5.1 describes two methods used to derive 2DLDs: 1) joint analysis, which is the commonly used method for deriving 2DLDs, and 2) combined analysis, which combines spectral linear discriminants (SLDs) and temporal linear discriminants (TLDs). Section 5.2 shows that this condition implies a certain structure of the across-class and within-class covariances. In Section 5.3, we use discriminants from both analyses on speech recognition task and show that discriminants from combined analysis perform better than those obtained from joint analysis. The sub-optimal performance of joint analysis is attributed to insufficient training data for joint analysis in Section 5.4. Finally, Section 5.5, compares results from this study to results from the previous studies and presents conclusions.

5.1 Two-dimensional Linear Discriminants

In this section, we derive discriminants in joint time-feature space. The input features are blocks of spectral vectors. Figure 5.1 shows one such block. It contains 101 frames, each of 15 dimensional spectral energies. The block is labeled by the phone label of the center

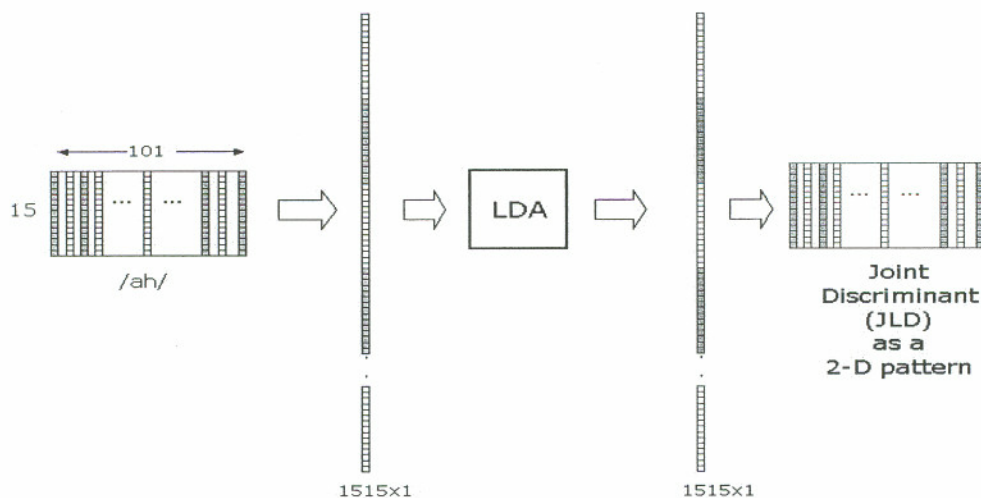


Figure 5.1: Procedure for obtaining JLDs.

frame. The center frame is also referred as current frame and the block contains 500 ms temporal information from the past and the future.

5.1.1 Joint Analysis

Joint analysis refers to LDA performed using the complete block of spectral energies. The features are the same as described in Chapter 2, and the development database is OGI Stories database. The process is described in Figure 5.1. First, the block of features is converted to a vector of $101 \times 15 = 1515$ dimensions. Each vector is labeled by the phone label of center frame. LDA is performed in 1515 dimensional feature space. Resulting LDs are 1515 dimensional vectors. Each vector is converted to 101×15 dimensional block and is interpreted as a two-dimensional pattern. This pattern is called a joint linear discriminant (JLD). Discriminant features are estimated by projecting blocks of spectral vectors on these patterns. We have used 24 joint discriminants in this work.

5.1.2 Combined Analysis

Combined analysis refers to a combination of spectral linear discriminants (SLDs) and temporal linear discriminants (TLDs). The resulting discriminants are called combined linear discriminants (CLDs). The analysis assumes that time and frequency domains are

independent. It means that only one set of SLDs and TLDs are estimated from the block of spectral vectors. All 101 spectral frames are projected on the same SLDs and all 15 temporal trajectories are projected on the same TLDs. In this work, SLDs are estimated using the center frame (or current frame) because it has the highest phone variability. We have investigated into different ways of combining statistics from 15 temporal trajectories: 1) using statistics from 5th band, 2) using statistics from 0th DCT coefficient trajectories, 3) using statistics from 0th SLD coefficient trajectories, and 4) estimating TLDs for each band separately and averaging them. We have observed that different TLDs do not change the recognition performance significantly but estimation of TLDs from 0th SLD coefficient trajectories gives the best performance (see Section 5.3). Further, it was observed that if the block of spectral energies is first filtered using any of the TLDs then the resulting SLDs from the center frame are same as those used in the estimation of TLDs.

In this work, CLDs are also derived using OGI Stories database. The features are the same as described in Chapter 2. The process of deriving CLDs is described in Figure 5.2. LDA in spectral domain is performed using 15 spectral energies as features. Each feature vector is labeled by the phone label of the corresponding speech segment. Spectral discriminant features are obtained by projecting spectral features on the SLDs. LDA in temporal domain is performed using a 101 dimensional temporal vector. Each vector is obtained by projecting a block of spectrogram on the first SLD, and it is labeled by the phone label of the center frame. Outer product of the resulting TLDs and SLDs gives CLDs. Note that CLDs can be interpreted as two-dimensional bases. If $\{\bar{e}_t|_{1 \times m}\}$ are TLDs and $\{\bar{e}_s|_{1 \times n}\}$ are SLDs then CLDs, $\{\bar{e}_{st}|_{m \times n}\}$, are

$$\begin{aligned} i &= 1, \dots, n_c - 1 \\ \bar{e}_{st}^k &= (\bar{e}_t^i)^t * \bar{e}_s^j, \quad j = 1, \dots, n_c - 1, \\ k &= 1, \dots, n_c - 1 \end{aligned} \quad (5.1)$$

where n_c is the number of classes. Subsequently, the process of applying the CLDs can also be considered as projecting the $m \times n$ block of spectrogram on these discriminants.

The optimal combination of SLDs and TLDs is decided by the criterion that is described in the next section. The results show that combination of first 8 SLDs and first 3 TLDs gives the optimal 24 CLDs. They are shown in Figures 5.3 and 5.4 respectively.

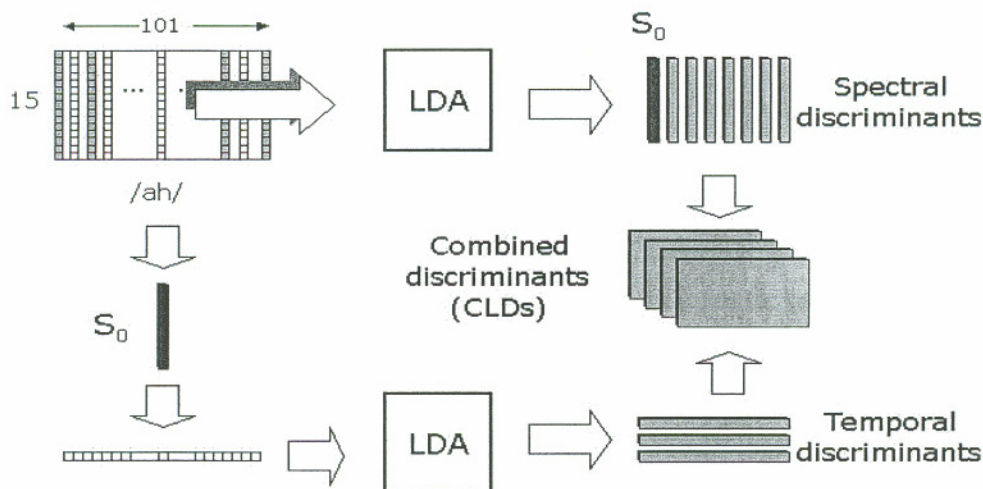


Figure 5.2: Procedure for obtaining CLDs.

Figure 5.5 shows four patterns obtained by combining the first two SLDs and TLDs. SLDs are same as the ones used in [26]. TLDs are same as the ones used in [51]. They attenuate the DC component and frequency components higher than 20 Hz. First and third TLDs are approximately symmetric filters and second TLD is approximately anti-symmetric filter. Note that second and third TLDs can also be approximated by the derivative and the double-derivative of first TLD.

Combined analysis breaks the joint analysis in two parts. It assumes that time and frequency domains can be optimized separately for the optimal discriminants in joint time-frequency domain. It ignores time-frequency correlations which may be important for speech recognition task. However, each stage of estimation uses fewer features than the joint analysis, and needs order of magnitude less data for estimating the discriminants.

5.2 Optimality of Combined Discriminants

The derivation of combined discriminants makes an assumption that spectral and temporal domains are independent. In the joint time-frequency domains, this assumption means that the variance of a feature element at time t and at frequency s is

$$\sigma_{st}^2 = \sigma_s^2 * \sigma_t^2. \quad (5.2)$$

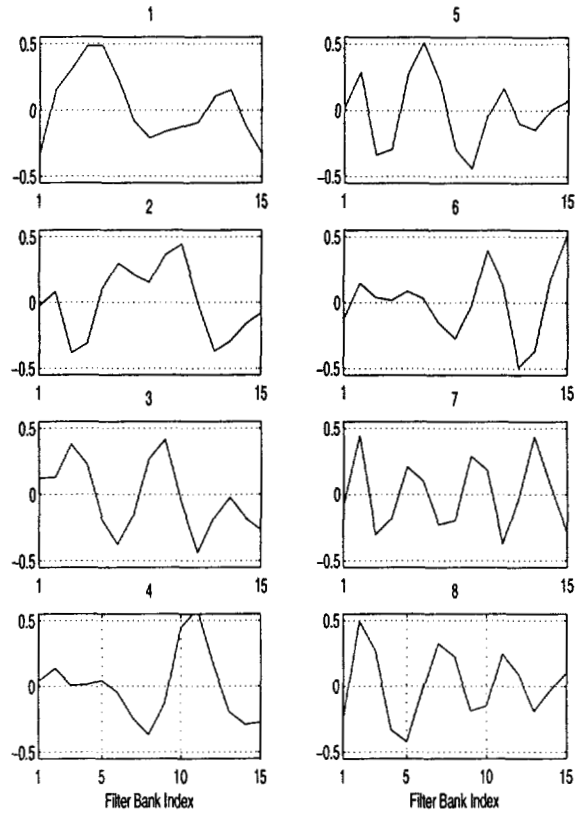


Figure 5.3: The First 8 LDs from spectral domain

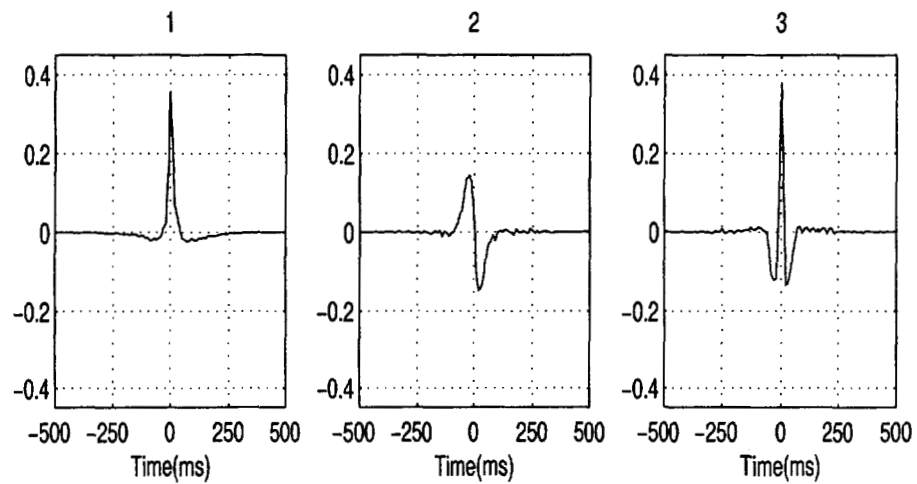


Figure 5.4: The 3 LDs from temporal trajectory of first spectral discriminant feature

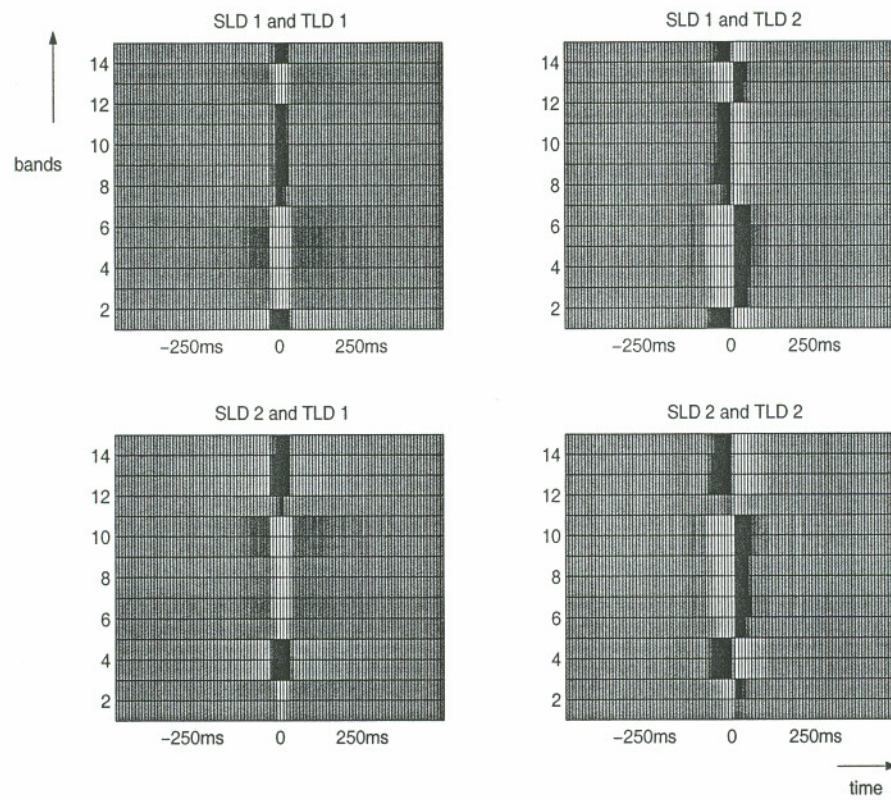


Figure 5.5: CLDs obtained by combining first two SLDs and TLDs. The title indicates the index of the SLD and TLD.

Therefore the covariance matrix estimated from different bands (s_i) over the complete time window $T = \{t\}$ is

$$\Sigma_{s_i T} = \sigma_{s_i}^2 * \Sigma_T, \quad (5.3)$$

and the cross-covariance matrix between two bands (s_1, s_2) over T is,

$$\Sigma_{s_1 s_2 T} = \sigma_{s_1 s_2} * \Sigma_T. \quad (5.4)$$

Thus the total covariance matrix has a special structure,

$$\Sigma_{ST} = \begin{bmatrix} \sigma_{s_1}^2 \Sigma_T & \sigma_{s_1 s_2} \Sigma_T & \cdots & \sigma_{s_1 s_{15}} \Sigma_T \\ \sigma_{s_2 s_1} \Sigma_T & \sigma_{s_2}^2 \Sigma_T & \cdots & \sigma_{s_2 s_{15}} \Sigma_T \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{s_{15} s_1} \Sigma_T & \sigma_{s_{15} s_2} \Sigma_T & \cdots & \sigma_{s_{15}}^2 \Sigma_T \end{bmatrix}.$$

In this case, the across-class and within-class covariances too have the same structure. Let W_T and W_S be the within-class covariances in time and spectrum. Let A_T and A_S be the across-class covariances in time and frequency. The across-class and within-class covariances in joint time-frequency domains will be,

$$A_{ST} = \begin{bmatrix} a_{s_1}^2 A_T & a_{s_1 s_2} A_T & \cdots & a_{s_1 s_{15}} A_T \\ a_{s_2 s_1} A_T & a_{s_2}^2 A_T & \cdots & a_{s_2 s_{15}} A_T \\ \vdots & \vdots & \ddots & \vdots \\ a_{s_{15} s_1} A_T & a_{s_{15} s_2} A_T & \cdots & a_{s_{15}}^2 A_T \end{bmatrix} \quad (5.5)$$

$$W_{ST} = \begin{bmatrix} w_{s_1}^2 W_T & w_{s_1 s_2} W_T & \cdots & w_{s_1 s_{15}} W_T \\ w_{s_2 s_1} W_T & w_{s_2}^2 W_T & \cdots & w_{s_2 s_{15}} W_T \\ \vdots & \vdots & \ddots & \vdots \\ w_{s_{15} s_1} W_T & w_{s_{15} s_2} W_T & \cdots & w_{s_{15}}^2 W_T \end{bmatrix}$$

Under these conditions, the joint linear discriminants have a special form. Let $\{\bar{e}_t\}$ and $\{\bar{e}_s\}$ be the set of eigenvectors in time and frequency domains. It can be shown that the optimal joint linear discriminants, $\{\bar{e}_{st}\}$, are,

$$\bar{e}_{st}^k = \begin{cases} \begin{bmatrix} \bar{e}_t^i & 0 & \cdots & 0 \\ 0 & \bar{e}_t^i & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \bar{e}_t^i \end{bmatrix} * \bar{e}_s^j, & \begin{matrix} i = 1, \dots, n_c - 1 \\ j = 1, \dots, n_c - 1 \\ k = 1, \dots, n_c - 1 \end{matrix} \end{cases}, \quad (5.6)$$

where n_c is the number of classes. The proof is given in appendix. It uses the fact that \bar{e}_{st}^k must simultaneously diagonalize A_{ST} and W_{ST} , that is,

$$\begin{aligned} A_{ST}\bar{e}_{st}^k &= d_{st}^k \bar{e}_{st}^k \\ W_{ST}\bar{e}_{st}^k &= \bar{e}_{st}^k, \end{aligned} \quad (5.7)$$

where, d_{st}^k is a scalar.

The number of combined LDs generated using the equation 5.6 are $m \times n$, if $m \times n \leq (n_c - 1)^2$; and $(n_c - 1)^2$, if $(n_c - 1)^2 \leq m \times n$. However, the optimal number of joint discriminants can only be $n_c - 1$ (assuming $n_c - 1 \leq m \times n$). The dimensionality of the feature space is $m \times n$, and there can be only $m \times n$ linearly independent vectors. So, there are two cases: 1) $m \times n < (n_c - 1)^2$, and 2) $(n_c - 1)^2 \leq m \times n$. In case 1, $m \times n$ vectors can be divided into 2 sets: I) $n_c - 1$ linearly independent joint discriminants; and II) $m \times n - n_c + 1$ linearly independent vectors which are orthogonal to the subspace spanned by joint discriminants. In case 2, $(n_c - 1)^2$ vectors can be divided into 2 sets: I) $n_c - 1$ linearly independent joint discriminants; and II) $n_c^2 - n_c + 2$ linearly independent vectors which are orthogonal to the space spanned by joint discriminants.

The choice of $n_c - 1$ leading discriminant vectors is based on the following criterion,

$$FR^k = \frac{(\bar{e}_{st}^k)^T A_{st} \bar{e}_{st}^k}{(\bar{e}_{st}^k)^T W_{st} \bar{e}_{st}^k}. \quad (5.8)$$

For each discriminant vector, FR^k is computed and they are ordered using FR^k . First $(n_c - 1)$ vectors are chosen as the optimal discriminants.

5.3 2D Discriminants in Speech

LDs are evaluated on a continuous digit recognition task using an independent database - OGI NUMBERS Database. The vocabulary is 11 words (0-9 and "oh"). In these experiments, each word is modeled as a sequence of context-independent monophones and each monophone is modeled using five-state, three-component hidden Markov model (HMM). Baseline system contains 8 coefficients obtained by projecting the spectral features onto discrete cosine transform (DCT) bases. The resulting features are appended with 8 delta

Table 5.1: Connected digit recognition performance using 24 discriminant features.

| Temporal Context | JLDs | CLDs |
|------------------|------|------|
| w=9 frames | 7.5 | 6.3 |
| w=101 frames | 6.9 | 5.3 |

and 8 double-delta coefficients computed using 4 and 9 frames respectively. The 24 features are finally processed using UMS. Their performance is 6.3 % word error rate (WER).

Table 5.1 shows the results of the speech recognition experiments. It shows that the longer temporal context improves recognition performance. Both CLDs and JLDs from a block of 100 frames perform significantly better than those obtained from a block of 9 frames. Results also show that CLDs always outperform JLDs. With temporal context of 9 frames, CLDs perform same as baseline features. With wider context, they perform significantly better than both JLDs and baseline features.

Table 5.2 shows the results for different types of CLDs. In all the cases, SLDs are estimated from the center frame or current frame. TLDs are estimated using four different methods: 1) using temporal trajectories from 5th critical band, 2) averaging filters from 15 bands, 3) using temporal trajectories of 0th cepstral coefficient, and 4) using temporal trajectories of first SLD coefficient. Cases 1, 3 and 4 estimate only one set of TLDs and use it on all the trajectories. The choice of 5th band, 0th cepstral coefficient, or first SLD coefficient is based on the fact that they have the highest phone variability among the other streams. Case 2 estimates filters from all the bands. These filters are averaged over different bands to obtain one set of filter that are used on all the bands. The difference between cases 2 and 3 is that the TLDs are averaged over different bands in the former case, and trajectories are averaged over different bands in the later case. The result shows that there is not significant difference in the performance but TLDs from first SLD coefficient stream performs slightly better than others.

5.4 Data Insufficiency for Joint Discriminants

In the previous section, we showed that CLDs outperform JLDs on continuous digit recognition task. This result is interesting because CLDs are a special case of JLDs, and given

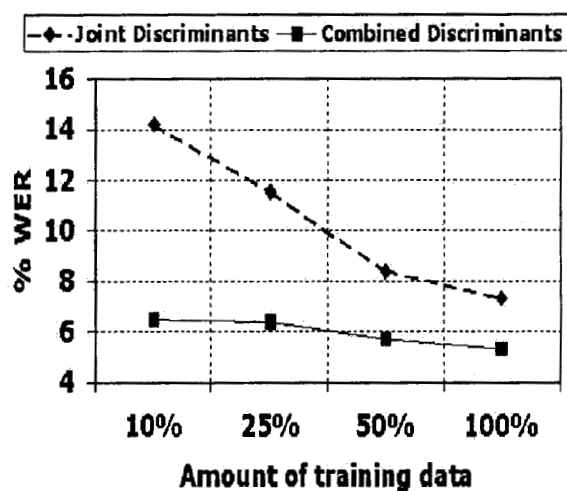
Table 5.2: Effect of different types of TLDs on CLDs

| TLDs derived from | %WER |
|--------------------------------|------|
| 5th critical band | 5.5 |
| Averaged filter over all bands | 5.6 |
| C0 | 5.5 |
| S0 | 5.3 |

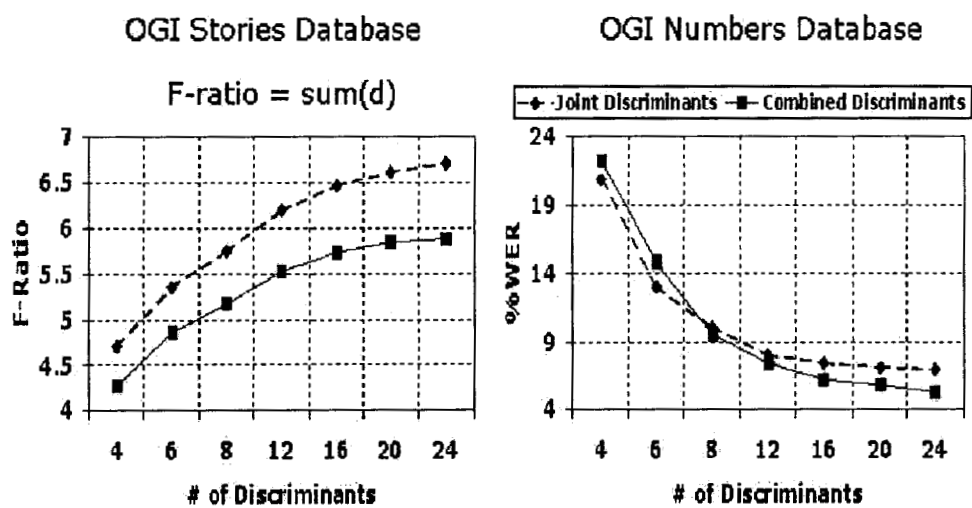
sufficient amount of data CLDs cannot outperform JLDs. Our hypothesis is that JLDs perform worse than CLDs because of insufficient data to estimate them. We prove this hypothesis using two experiments. First we reduce the data used for estimating the LDs while keeping the input feature dimensionality same. Second we compare F-ratio on the training database with recognition performance on the evaluation database.

Figure shows the results of the first experiment. We have used LDs estimated from 1/10th, 1/4th, 1/2 and complete database. In each case, across-class and within-class covariances were estimated by averaging the same statistics from 10 subsets of the original database. Results show that performance of JLDs degrades much faster than CLDs. Using 1/4th of the data size, WER for JLDs is almost twice that of using the complete database. With the same difference, difference in WER for CLDs is only about 1%. This shows that JLDs need more training data than CLDs and JLDs will perform similar to CLDs with larger training data.

Figure 5.6 shows the results of the second experiment. In this case, we compare two results - cumulative F-ratio for discriminants on the training database, and recognition performance using them on evaluation database. F-ratio is obtained by adding the eigenvalues corresponding to the eigen vectors. Since it is a measure of discriminability, higher F-ratio implies better phone separability. Results show that JLDs always have higher F-ratio than CLDs on the training database. On evaluation database, first few JLDs perform better than CLDs. However, higher JLDs do not give significant advantage over the corresponding CLDs. This shows that the higher JLDs are not generalizing on an unseen task. In other words, JLDs over-fit the training data when sufficient training data is not available.



(a) Effect of datasize on the estimation of LDs



(b) Performance of LDs on development and evaluation databases

Figure 5.6: Insufficient training data for JLDs. (a) shows that JLDs need more training data than CLDs. (b) shows that in absence of sufficient training data JLDs over-fit the training data.

5.5 Conclusions

In this chapter, we derived joint time-frequency discriminants using a longer time span than 9 frames. This was based on an earlier observation that information about current phone lasts beyond its boundaries. A block of 101 frames of spectral vectors was used as features. LDA was used to derive a small set of features that contains the most useful variability in this block. Two methods of applying LDA were studied. First method, also the most commonly used method, used the complete block as a feature vector. It was called joint analysis. A new method was proposed, that combined SLDs and TLDs to obtain two-dimensional discriminants. It was called combined analysis. The discriminant features from both analysis were used in continuous digit recognition task. The results showed that discriminant features from combined analysis perform better than those from joint analysis.

Worse performance of the joint analysis was attributed to insufficient training data. This hypothesis was proved using two experiments. Results of first experiment showed that performance of discriminants from joint analysis degrades faster than those from combined analysis when the amount of training data is reduced. Results of second experiment showed that joint analysis over-fits the training data when sufficient data is not available. Therefore, the discriminant features from it give consistent improvement on the development database but not on the evaluation database.

Bahl, et. al. [35] have obtained a similar result on a large vocabulary continuous speech recognition task. Our study extended their result in many ways. First, we observed that TLDs need not be derived from different spectral discriminant streams. Better performance was obtained using the TLDs from the first spectral discriminant stream on all the other streams. This was attributed to the fact that higher spectral discriminant streams have higher signal-to-noise ratio and the TLDs derived from them are noisier. Second, we observed that longer temporal discriminants (≈ 1000 ms) outperform the discriminants derived from the shorter time span of 90 ms.

Chapter 6

Summary and Conclusions

The thesis proposed analysis of variability of the features. The results of the analysis were used in to prove two hypotheses: 1) variability in the feature-set is related to its performance on speech and speaker recognition tasks, and 2) results of the analysis can be used to derive new feature set that is robust to the mismatch in training and testing conditions.

6.1 Summary

The thesis was organized in three parts. Chapter 2 is the first part which described the statistical analysis technique MANOVA. Using this technique, variability in the most commonly used features was decomposed in spectral and temporal domains. It was shown that, in spectral domain, phone variability is highest in 4-6 Barks. In temporal domain, phone variability is 250 ms around the current frame. This showed that effect of variation at current phone lasts beyond its boundaries and into the neighboring phones.

The features were also analyzed with and without UMS – a commonly used preprocessing technique in speech and speaker recognition systems. In both domains, it was also observed that channel variability is higher than speaker variability before UMS. This showed that mean of the utterance is more sensitive to channel variations than speaker variations. After UMS, features had more speaker variability than channel variability. Since this variability is the interaction between phones and speakers and channels, we concluded that features within different phones are more sensitive to speaker variations than channel variations.

The results of the analysis were shown to be invariant to different data-sets and different databases. The results of the analysis were compared with the results of two previously proposed methods [15, 23]. In the first case, similar features, as used in [15], were analyzed on TIMIT database. The results from the proposed analysis were similar to the results obtained in [15]. In the second case, the results of MANOVA were converted to mutual information and they were compared to the results of [23]. Again, it was observed that the results from both analysis were similar.

In the second part of the thesis, which includes Chapter 3, variability in the features was related to the performance of features on speech and speaker recognition tasks. Speaker and channel variability within four broad phonetic categories – vowels+diphthongs, glides+nasals, fricatives, and silence+stops – was measured using the proposed analysis. This was compared to the speaker recognition performance obtained using the features corresponding to the phone categories. It was shown that features from vowels+diphthongs are the most sensitive features to speaker variations and they also perform the best on the speaker recognition task. This is followed by features from glides+nasals and fricatives. The features for silence+stops were the least sensitive features for speaker variations and they performed worst on the speaker recognition task.

The relationship between phone variability and speech recognition system was shown in two experiments. First, the effect of appending delta and double-delta features was measured using MANOVA and continuous digit recognition experiments. The results showed that phone variability improves by appending delta and double-delta features. The word accuracy also showed improvement by appending them. Second, the effect of different feature transformations was analyzed using MANOVA. These features were evaluated on AURORA task. The comparison showed that the transformations improved phone variability in the features and also improved word accuracies on AURORA task.

In the third part, which includes chapters 4 and 5, described how results of MANOVA can be used to derive robust features for speech recognition task. In Chapter 4, we grouped the results of MANOVA into two types: useful variability and harmful variability for speech recognition task. It was assumed that phone variability is the useful variability and the remaining variability is the harmful variability. For LDA, the useful variability was used

as across-class covariance and harmful variability was used as within-class covariance. The resulting bases were called as linear discriminants (LDs). The new feature-set was obtained by projecting original features on linear discriminants and they were called discriminant features.

In temporal domain, we investigated two ways of modeling stops as classes: 1) burst and closure modeled as one class and 2) burst and closure modeled as two classes. Since closure is followed by burst, modeling them separately implied modeling two classes that will mostly follow each other. It was observed that when they are modeled as two classes, one of the LDs is a delta-like filter. This result is significant because the delta filter was derived from the data, and it is shown to discriminate between the classes that follow one another.

This experiment was extended by modeling all the phones as a sequence of sub-phones. Each subphone was analogous to modeling each phone using multi-state HMM and considering each state as a subphone. This was motivated by the fact that there was difference in phone models used in MANOVA, and the ones used in speech recognition experiments. MANOVA had assumed that phone is modeled as a one-state one-component HMM, but phones are usually modeled as multi-state multi-component HMMs. To make the analysis similar to the recognition setup, LDA was performed using sub-phones as classes.

Resulting LDs also showed a delta-like filter. It was due to the fact that some of the sub-phones always followed one another. We compared the performance of these LDs with those obtained using phones as classes on digit recognition task. The LDs using sub-phones as classes outperformed the LDs using phones as classes.

In Chapter 5, we used results of MANOVA to derive LDs in joint time-frequency domain. This was based on an earlier conclusion that variability due to current phone lasts for about 250 ms around it. Therefore, the goal was to incorporate longer time-span in the feature extraction. A trivial approach is to use a wider block of spectral vectors directly as features. This results in high dimensional features that are not suitable for statistical modeling due to high correlation between the feature elements and due to curse of dimensionality. In this work, LDA was used to derive a smaller feature-set that contains most of the useful variability for speech recognition. The resulting discriminants

were called two-dimensional discriminants.

LDA was performed using two methods. First all the correlations in the block of spectral vectors were retained. This was referred to as joint analysis. Second, time and frequency domains were assumed to be independent. LDA was performed independently in time and frequency domains, and the discriminants were combined to form two-dimensional discriminants. This method ignored time-frequency correlations but used a smaller set of features for the analysis. The analysis is referred to as combined analysis. Note that this was a special case of joint analysis.

Discriminants using the two methods were evaluated on speech recognition task. Results showed that features from the combined analysis outperform those from the joint analysis. Further analysis showed that joint analysis needs more training data than combined analysis. In absence of that, discriminants from joint analysis over-fit training data, and have poor generalization on the evaluation data.

6.2 Future Directions

This work presented analysis of variability in features assuming that distribution of feature is multivariate Gaussian distribution. This work can be extended in many ways. First the analysis can be modified to model the data using a mixture of Gaussians. This will lead to an accurate estimates of the probability density functions and an accurate estimate of different variabilities in speech.

The conventional feature processing steps such as short-time Fourier transform, critical-band warping, non-linear compression can be re-evaluated using the proposed technique for speech and speaker recognition. Conventionally, Bark or Mel warping, and logarithmic nonlinearity are used in the features. Their importance has been shown in Chapter 1. The re-evaluation refers to modification of the procedure and its evaluation using MANOVA.

MANOVA can be also used to investigate the residual features that are ignored in the conventional feature extraction methods. These include phase of the signal and the error in approximating 128 dimensional features using 15 dimensional Bark spectrum.

In Chapter 3, it was shown that vowels and diphthongs are more important than others.

The nature of speaker variability in those classes needs to be investigated. In phone-based speaker recognition, the speaker recognition accuracy depends on both phone recognition accuracy and speaker recognition accuracy within each phone. This relationship needs to be explored when deriving new features for speaker recognition.

Finally, it is shown in appendix E, the variability in speech can be quantified using different formulas. The requirement for a good formulation is that

- it should increase with increase in the number of features,
- the variability in the sources should add up to the total variability, and
- it should have meaningful interpretation.

There is only one measure that satisfies these requirements and it is mutual information (MI). It increases by adding more features. The information in the sources adds up to the total information in the features. It also has more meaningful interpretations than % contribution of variabilities. Usability of this measure needs to be investigated further.

6.3 Final Comments

MANOVA proposed for analyzing variability in features has some limitations. Since it assume that testing data are the same as training data, it can not be used to study the effects of generalization. For example, Chapter 5 showed that joint analysis always gives higher F-ratio than combined analysis. This was due to over-fitting of the training data. Joint analysis performed worse than combined analysis on recognition tasks. Therefore, MANOVA must be performed on an independent and large database.

Through some examples in Chapter 3 it was shown that increase in phone variability leads to increase in the speech recognition performance. In general, speech recognition performance depends on many factors including the effects of language model. MANOVA ignores the effect of language model, it assumes that any phone can follow any other phone. Our results on digit recognition tasks, which have a simple language model, have shown correlation between phone variability and recognition accuracy. This correlations needs to verified on large vocabulary tasks.

Finally, it is shown in the thesis that variability in the features is related to the performance of features on speech and speaker recognition tasks. It was also shown that MANOVA can be used to derive discriminant features for speech recognition tasks. The proposed MANOVA is computationally inexpensive and it has a faster execution than running speech and speaker recognition experiments. We hope that it will be used in future research both for deriving better features and for increasing our understanding of variabilities in speech signal.

Appendix A

Phone Set for MANOVA

Table A.1: Phone set common between OGI Stories, HTIMIT and TIMIT databases.

| Phone | Example | Phone | Example | Phone | Example | Phone | Example |
|-------|-----------------|-------|-----------------|-------|----------------|-------|-----------------|
| aa | f <u>a</u> ther | dx | wri <u>t</u> er | jh | ju <u>d</u> ge | s | si <u>g</u> n |
| ae | b <u>a</u> t | eh | b <u>e</u> t | k | <u>c</u> an | sh | ass <u>u</u> re |
| ah | ab <u>o</u> ve | er | bi <u>r</u> d | l | <u>l</u> ent | t | <u>t</u> an |
| ao | ca <u>u</u> ght | ey | ba <u>y</u> | m | <u>m</u> e | th | <u>t</u> hig |
| aw | ab <u>o</u> ut | f | <u>f</u> ine | n | <u>k</u> nee | uh | bo <u>o</u> k |
| ay | by <u>e</u> | g | ga <u>n</u> der | ng | si <u>n</u> g | uw | bo <u>o</u> t |
| b | <u>b</u> an | h | <u>h</u> ope | ow | bo <u>o</u> t | v | <u>v</u> ine |
| ch | <u>ch</u> urch | ih | bi <u>t</u> | p | <u>p</u> an | z | res <u>ig</u> n |
| d | <u>d</u> an | iy | be <u>e</u> t | r | <u>r</u> ent | | |

Appendix B

Effect of Phonetic Context on Results of MANOVA

MANOVA is performed in spectral and temporal domains using context-dependent monophones as classes. The phonetic context is specified in terms of 4 broad phonetic categories - vowels+diphthongs, glides+nasals, fricatives, silence+stops. Figures B.1 and B.2 shows the nature of variability in the same features used in chapter 2 with and without explicit context modeling. Table B.1 shows contribution of variabilities for the two cases.

The results show that context variability is a significant part of residual variability. Therefore the contribution of residual variability reduces when context-dependent phones are used as classes for MANOVA. The new phone variability is higher than the phone variability computed using phones. The results also show increased interaction between speakers and the context-dependent phones and speakers.

In spectral domain, using phones as classes, the phone variability was dominant between 3-6 Barks and residual variability was dominant between 10-15 Barks. The results with context-dependent phones show that the new phone variability is dominant over all frequency components. In temporal domain, using phones as classes, effect of the phone variability was shown to spread for approximately 250 ms around the current frame. Using context-dependent phones as classes, the effect spreads for longer than 500 ms around the current frames. This is due to the increase in the length of the phone unit, with context-dependent phones we are modeling three phones as one unit.

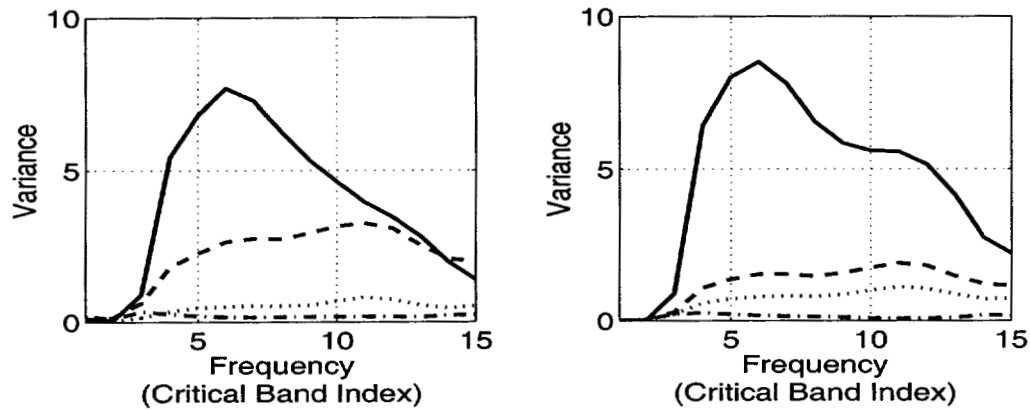


Figure B.1: Results of MANOVA using context-independent phones and context-dependent phones as classes in spectral domain

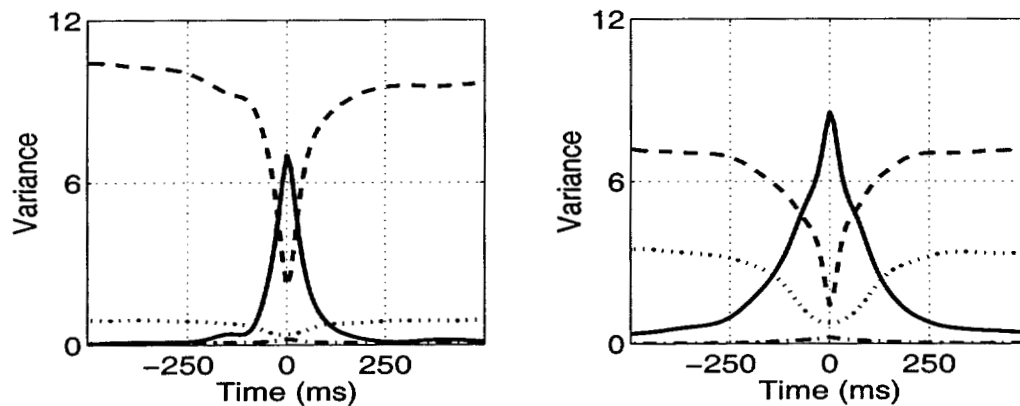


Figure B.2: Results of MANOVA using context-independent and context-dependent phones as classes in temporal domain

Table B.1: Contribution of variabilities using context-independent (CI) phones and context-dependent (CD) phones as classes

| | Spectral Domain | | Temporal Domain | |
|----------|-----------------|------|-----------------|------|
| | CI | CD | CI | CD |
| Phone | 58.1 | 69.5 | 7.0 | 17.5 |
| Speaker | 6.9 | 10.3 | 7.2 | 24.5 |
| Channel | 2.9 | 2.0 | 0.8 | 1.4 |
| Residual | 31.9 | 18.2 | 82.6 | 56.4 |

Appendix C

Temporal Variability in Different Bands

In chapter 2, we presented results in temporal domain from energies from fifth filter on Bark scale. In this appendix, we present results for energies from all the filters on Bark scale (figures C.1 and C.2). Note that, for each band, all the feature elements are analyzed independently. The results show that the nature of variability across different bands is similar. That is in all bands, the phone variability is highest at the current frame, residual variability is lowest at the current frame, and speaker and channel variabilities are almost constant. The results also show that phone variability in all bands, except first two bands, spreads for approximately 250 ms around the current phone. Note that the variance at the current frame (0 ms) in each band is the same as the variance of that frequency component in spectral domain.

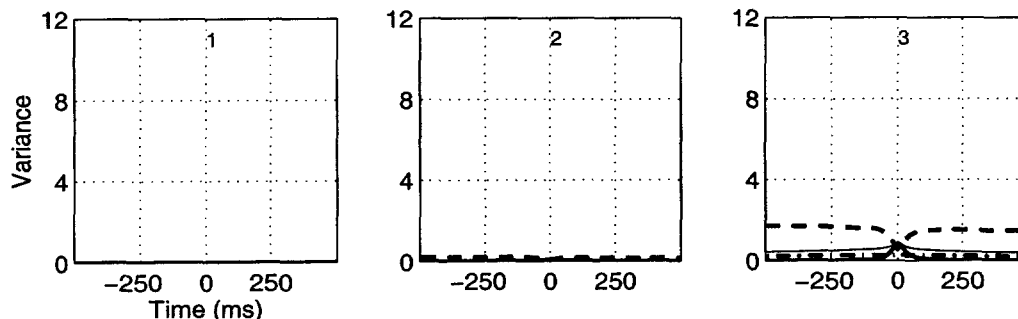


Figure C.1: Temporal variability in bands 1-3: phone variability (thick solid line), speaker variability (thin solid line), channel variability (thick dot-dash line), residual variability (thick dotted line). Number in each plot is the critical band index.

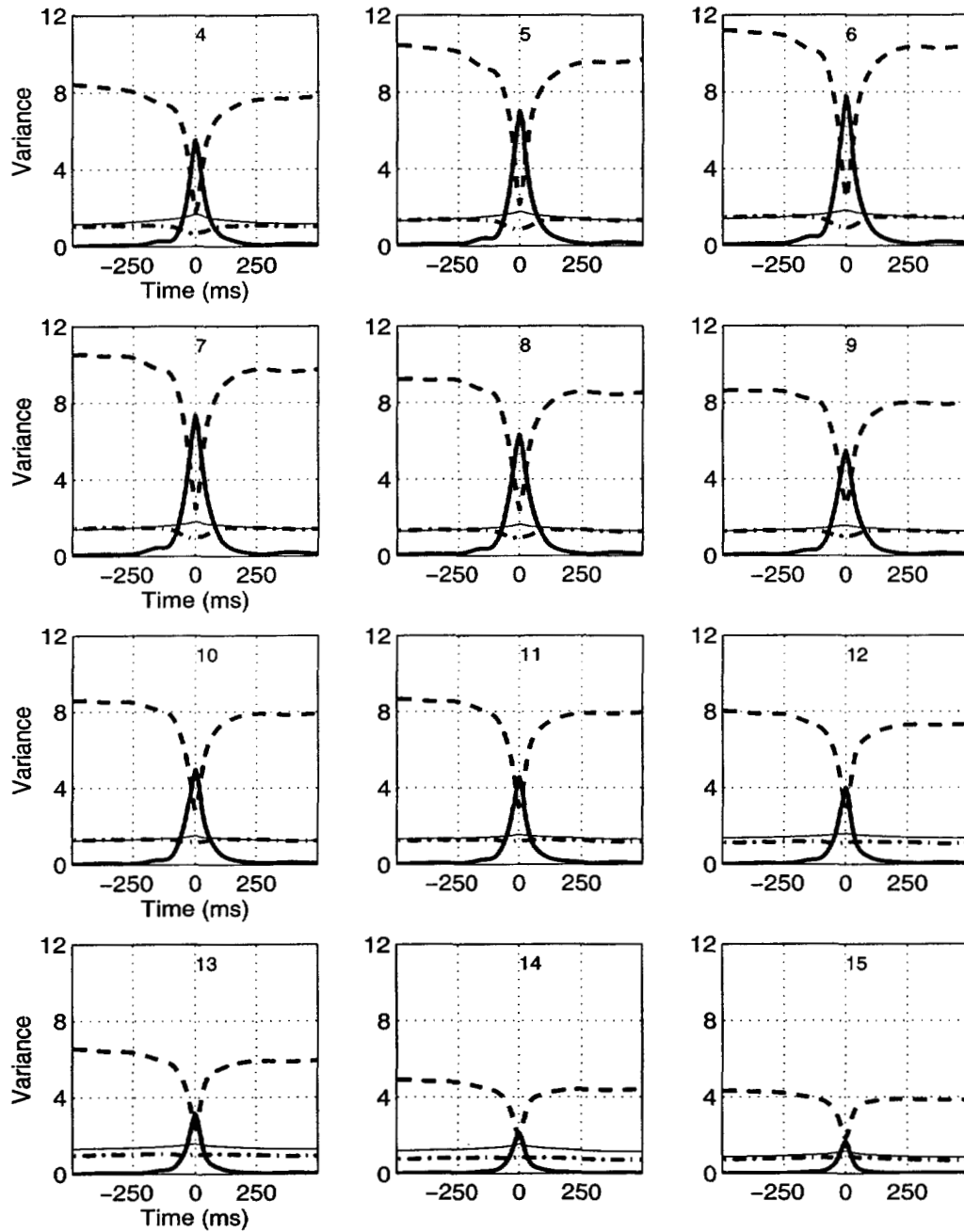


Figure C.2: Temporal variability in bands 4-15: phone variability (thick solid line), speaker variability (thin solid line), channel variability (thick dot-dash line), residual variability (thick dotted line). Number in each plot is the critical band index.

Appendix D

Temporal Filter for Speaker Verification

In this appendix, we design a temporal filter for speaker verification using results of MANOVA. For speaker recognition, we assume that phone+speaker variability is the useful variability and channel+residual variability is the harmful variability. The useful variability is phone+speaker variability instead of speaker variability because the features used for deriving the filter are preprocessed using UMS. Resulting speaker variability is the variability due to interaction between speakers and phones. If speaker variability is a useful variability then that also implies that phone variability is a useful variability too.

The features are same as the ones used in Chapter 2. Phone, speaker+channel, and residual variabilities are estimated using stories database as described in Chapter 2. Channel variability is estimated using HTIMIT database using the procedure illustrated in [43]. All the statistics are estimated from fifth critical band. LDA is used to derive the filter. The across-class covariance (A_c) is an estimate of phone+speaker variability and within-class covariance (W_c) is an estimate of channel+residual variability. The filter is the leading eigen vector obtained using $W_c^{-1}A_c$. The discriminant features are obtained by filtering the time trajectories of spectral energies from all critical bands using this filter.

Figure D.1 shows the impulse and frequency response of the resulting filter. The resulting filter is an approximately symmetric band-pass filter which attenuates frequency components below 1 Hz and above 15 Hz. This is filter is similar to the data-driven RASTA filter (see Chapter 4 and [56]) used in speech recognition. It is also similar to the symmetric RASTA [27].

The filter is evaluated on 2001 NIST cellular speaker verification task. The system is implemented in UBM-GMM framework. It has 256 component GMM that is modeled

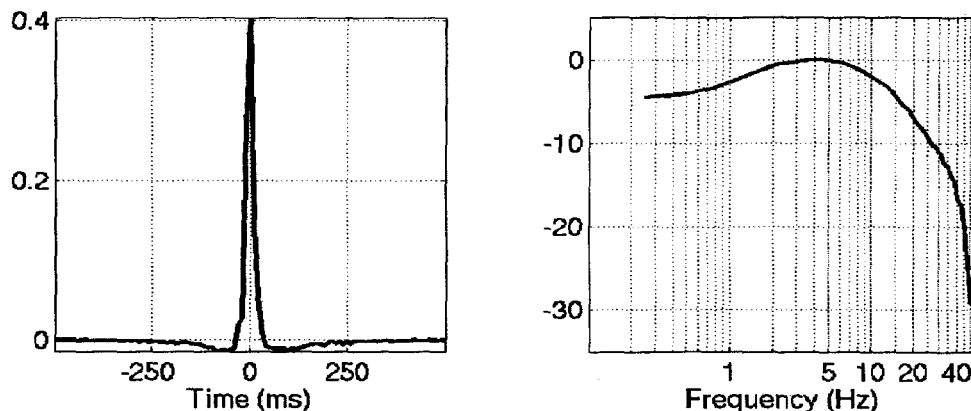


Figure D.1: Impulse and frequency response of the filter used in speaker recognition system using 39 component features. The features are 13 MFCCs filtered using the proposed filter and are appended with 13 delta and 13 double delta features. Each feature stream is processed using UMS. High energy frames are detected using adaptive energy-based speech-sil segmenter, and are used in these experiments.

Performance of this filter is compared to the baseline system that uses only UMS and to the system that uses a previously proposed data-driven filter [43]. Difference between the two filters is that the former filter is estimated using phone+speaker variability as useful variability and channel+residual variability as harmful variability, and the later filter was estimated using phone variability as useful variability and channel variability as harmful variability.

Figure D.2 shows the results obtained using the three configurations. System with UMS has 10.2% EER. System with the filter proposed in [43] has 10.4% EER. System with the proposed filter has 9.2% EER. Results show that the new filter improves the performance of the system over the system that uses only UMS. The improvement in the performance compared to the previously proposed filter shows the importance of using speaker and residual variability in the design of the filter

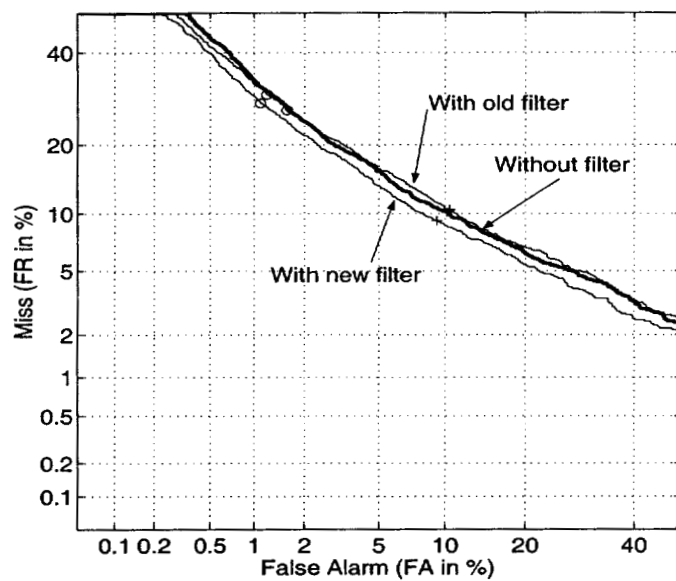


Figure D.2: Performance of old and new filter on speaker verification

Appendix E

Different Measures for Source Contributions

In Chapter 3, the relative phone variability is estimated using F-ratio assuming phones as classes. In this appendix, we compare different measures that can be used to estimate relative phone variability. These measures are computed using results of MANOVA for four different feature sets. The measures are correlated with the speech recognition performance obtained using these feature-sets. The feature-sets are (1) logarithmic filterbank energies (or same number of DCT coefficients from them), (2) 12 DCT coefficients + 12 Δ coefficients, (3) 8 DCT + 8 Δ + 8 $\Delta\Delta$ coefficients, and (4) 8 DCT + 8 Δ + 8 $\Delta\Delta$ coefficients with cepstral mean subtraction (CMS). Note that all the feature-sets contain 24 features.

The relative phone variability in the feature-sets is computed using HTIMIT database as follows. First, total variability is decomposed as,

$$\Sigma_{total} = \Sigma_{phone} + \Sigma_{speaker} + \Sigma_{channel} + \Sigma_{residual}$$

,where Σ_{total} is the total covariance or a measure of total variability, Σ_{phone} is the intra-phone covariance or a measure of phone variability, $\Sigma_{speaker}$ is the intra-speaker covariance or a measure of speaker variability, and $\Sigma_{channel}$ is the intra-channel covariance or a measure of channel variability. $\Sigma_{residual}$ is the covariance due to unaccounted sources or a measure of residual variability. Then, the relative phone variability is measured using 5 different measures -

- $\%PHN1(phone) = trace(\Sigma_{phone})/trace(\Sigma_{total})$. This is what I have been using in

the thesis. This is invariant to orthonormal linear transformation. For example, %PHN1 is same before and after discrete cosine transformation (DCT). But it is not same after whitening. After whitening it becomes %PHN2. This measure may not increase by adding more features in the feature-set.

- $FR1(phone) = trace((\Sigma_{total} - \Sigma_{phone})^{-1}\Sigma_{phone})$. $\Sigma_{total} - \Sigma_{phone}$ is the within-phone covariance and Σ_{phone} is the across-phone variance. This measure is also referred as F-ratio. This is invariant to any linear transformation. This measure increases when the number of features in the set are increased. But, it does not satisfy the condition

$$FR1(phone) + FR1(speaker) + FR1(channel) + FR1(residual) \neq 1$$

- $\%PHN2(phone) = trace(\Sigma_{total}^{-1}\Sigma_{phn})/24$. This a combination of %PHN1 and %FR1. The similarity between %PHN1 and %PHN2 is that all the source variabilities and the residual variability adds to 100 %. They are identical when $\Sigma_{total} = I_{24 \times 24}$, identity matrix. Difference between them is that %PHN1 does not consider orientation of the covariances in the feature space whereas %PHN2 does. Similarity between FR1 and %PHN2 is that they are correlated, that is, if one changes then the other changes in the same way. The difference is that FR1 over all source variabilities and residual variabilities does not add up to 100 %, but %PHN2 does. Finally, %PHN2 may not increase by adding more features in the feature-set.
- $FR2(phone) = trace(\Sigma_{residual}^{-1}\Sigma_{phone})$. This is similar to FR1 except that phone variability is measured with respect to the residual variability. This is called as Hotelling's Trace criterion and it is used in MANOVA literature to measure an effect of source. This is also invariant to linear transformation. This measure increases when the number of features in the set are increased. However, this measure does not satisfy this condition

$$FR2(phone) + FR2(speaker) + FR2(channel) + FR2(residual) = FR2(total) \neq 1$$

So, we modify FR2 to $\%FR3 = FR2(source)/FR2(total)$. Now it satisfies the above criterion.

- $MI(phone) = \log\left(\frac{|\Sigma_{total}|}{\prod_i |\Sigma_i| p_i}\right)$, where i denotes different phones and Σ_i is the covariance within each phone. This is an information-theoretic method for measuring the effect of a source. MI is measured in *nats*. This is invariant to linear transformations. It increases when the number of features in the set are increased.

The recognition experiments are performed using OGI Numbers database. We use 23 context-independent phones, where each phone is modeled as 5-state, 3-component HMM. The language model is built assuming that any digit can follow any other digit.

| Feature Set | %PHN1 | FR1 | %PHN2 | MI | FR2 | %FR3 | %WER |
|-------------|-------|-----|-------|-----|------|------|------|
| (1) | 50 | 3.9 | 9 | 1.6 | 8.6 | 20 | 11.8 |
| (2) | 49 | 4.1 | 13 | 2.1 | 10.5 | 23 | 8.0 |
| (3) | 50 | 4.3 | 14 | 2.2 | 11.0 | 24 | 6.9 |
| (4) | 61 | 5.3 | 15 | 3.2 | 10.9 | 27 | 6.3 |

Table E.1: Different measurements using MANOVA and corresponding recognition performance

Table E.1 shows the results using different measures of contribution of phone variability and the performance of feature-sets on recognition task. Here are some conclusions from these results

- Change in %PHN1 correlates with %WER before and after CMS. However, it does not increase by adding Δ and $\Delta\Delta$ coefficients. The reason is explained below in the discussion of %PHN2.
- Change in FR1 correlates with %WER. It is also guaranteed to improve when the number features are increased. The drawback of this measure is that the relative contribution all the sources does not add up to 1. In addition, it is difficult to interpret the measure, for example, it does not answer the question how much FR1 is needed for the error-free phone recognition.
- Change in %PHN2 also correlates with %WER. This measure is shows less phone variability than %PHN1. Comparison of %PHN1 and %PHN2 shows that changes in the orientation of the phone covariance, which are ignored by %PHN1, play important role in their performance on speech recognition task. This is explained in figure

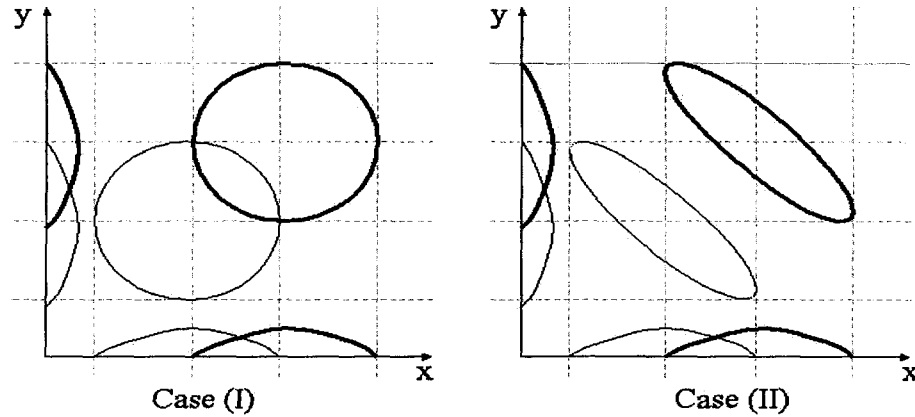


Figure E.1: Difference between %PHN1 and %PHN2. For Case (I) and Case (II), %PHN1 will be same because the marginal distributions are same. %PHN2 for these cases will be different because, in the joint space, classes have less overlap in Case (II) than Case (I).

E.1. Figure shows two cases where the marginal distribution of the classes is same but the joint distribution is different. Note that in both cases, %PHN1 will be same because $trace(\Sigma_{phone})$ and $trace(\Sigma_{total})$ are same. %PHN2 will, however, be higher in Case(II) than Case(I) because, in the joint space, the classes have less overlap in Case (II) than Case (I). Note that because of the normalization, this measure is not guaranteed to improve by increasing the number of features.

- Change in FR2 correlates with %WER in all cases except after CMS. Feature-set with CMS has a lower FR2 than the one without CMS. The reason is that CMS reduces $\Sigma_{speaker}$ and $\Sigma_{channel}$. FR2 does not account for changes in sources other than phone. This shows that the performance of feature-set also depends on the contribution of other source covariances apart from intra-phone covariance. %FR3 considers changes in all the sources and it correlates with %WERs for different features. However, it is not guaranteed to improve when the number of features are increased.
- Change in %MI also correlates with %WER. This is the best measure among all the others. It satisfies all the requirements. This measure is guaranteed to improve when the feature size improves. It satisfies the condition that MI over all sources adds up to the joint entropy of the phones, speakers, and channels. It also give

the highest bound for the phone information that can be computed from the phone priors. For HTIMIT, this bound is 4.53 nats.

Appendix F

Line Spectral Pair and MFC Coefficients

Recently, it was shown that line spectral pair (LSP) coefficients are suitable for the speaker segmentation task [6]. In this appendix, we investigate into the nature of variability in LSPs, and compare the results with similar results from Mel frequency cepstral coefficients (MFCCs).

Both features are computed from 32 ms speech segment and adjacent segments are shifted by 10 ms. LSPs are obtained by estimating 24 linear prediction coefficients and solving the roots of the polynomial [17]. MFCC coefficients are derived by projecting logarithmic energies from 24 filters on Mel scale onto 24 discrete cosine bases (C1-C24). Note that in both cases, the frame energy - C0 for MFCC and logE for LSPs - is not used. Note also that delta and double-delta features are not used in the following experiments. The choice of the features was based on their performance on the speaker segmentation task [6].

Table F.1 shows results of MANOVA for both features. We conclude that LSPs have higher phone and channel variability than MFCCs features whereas MFCCs have higher speaker variability. Therefore, we conclude that LSPs will perform better than MFCCs on speech recognition task and MFCCs will perform better than LSPs on speaker verification task. We verify these conclusions using speech and speaker recognition experiments.

Speech recognition experiments are performed on OGI Numbers database. It is continuous digit recognition with the vocabulary of 11 words (0-9 and "oh"). Each word is modeled as a sequence of context-independent monophones and each monophone is modeled as 5 state, 3 component HMM. Recognition results show that MFCCs give 26.3% WER and LSPs give 19.1% WER. This is in agreement with the results of MANOVA that

Table F.1: Comparison of source contributions using MANOVA using LSP and MFCC features.

| source | LSPs | MFCCs |
|----------|------|-------|
| phone | 36.9 | 13.7 |
| speaker | 10.9 | 17.5 |
| channel | 15.8 | 9.1 |
| Residual | 36.4 | 59.7 |

LSPs have the higher phone variability than MFCCs.

Speaker verification experiments are performed on 2001 NIST cellular speaker evaluation database. Verification was performed in GMM-UBM framework using 32 component GMM. The results show that MFCCs give 19.4 % EER and LSPs have 21.6 % EER. The higher EER for LSP coefficients is in agreement with the higher speaker variability in MFCCs.

The higher channel variability in LSPs is helpful in the speaker segmentation task as follows. Speaker segmentation is a task of identifying segments of individual speakers in a conversation. Typically speakers in a conversation also use different handsets. Therefore, speaker segmentation can be improved using features which are sensitive to handset variations. This is shown using LSPs and MFCCs on NIST 2001 two-speaker segmentation task. Refer to [6] for a detailed description of this task. It is observed that speaker segmentation error using LSPs is 6.0 % and the error using MFCCs is 10.0%. These results also confirm the results of MANOVA.

In this work, MFCCs were shown to comply with the assumptions of MANOVA, that is, additivity of the sources and normality of the distribution. LSPs, however, have not been examined for these conditions. LSPs loosely represent the locations of poles in the frequency plane. If the sources are modeled as a filter, then their effect in frequency plane can also be modeled as addition of poles in frequency domain. Therefore, the assumption behind the additivity holds for LSPs. However, LSP coefficients are highly correlated and the distribution of LSP coefficients is highly non-Gaussian. In this work, the uncorrelated features are obtained by projecting LSPs on the whitening transform [33]. However, the distribution of the resulting features is still highly non-Gaussian. Therefore,

more investigation needs to be done with these results to improve the estimates of different types of variabilities. However, the correspondence of the results of MANOVA with speech and speaker recognition results is very encouraging.

Appendix G

Analysis of Variability using Mutual Information

In Appendix E, we described different measures that can be used to quantify the variability due to sources. We showed that only one measure that is based on information-theoretic approach satisfies all the requirements for a good measure. Therefore, in this appendix, we propose an information-theoretic analysis of variability in speech. Similar to MANOVA, we assume that speech carries information from three main sources- language, speaker, and channel. We measure information from a source as mutual information (MI) [58] between the corresponding class labels and features. For example, linguistic information is measured as MI between phone labels and features. The effect of sources is measured in nats (or bits). In this work, we show it is easier to interpret the results of this analysis than the analysis of variability.

In general, MI between two random variables X and Y can be measured using three different methods [2]. First, assuming that X and Y have a joint Gaussian distribution. However, we cannot use this method because one of the variables - a set of class labels - is discrete. Second, modeling distribution of X or Y using parametric form, for example, mixture of Gaussians [2]. Third, using non-parametric techniques to estimate distributions of X and Y [30]. The proposed analysis is based on the second method, where distribution of features is modeled as a Gaussian distribution. Although it is a strong assumption, we show that results of this analysis are similar to the results obtained using the third method [30].

The paper is organized as follows. Section G.1 describes the experimental setup.

Section G.2 describes MANOVA and presents results of MANOVA. Section G.3 proposes information theoretic approach for analysis of information in speech and presents the results. Section G.4 compares these results with results from the previous study. Section G.5 describes the summary and conclusions from this work.

G.1 Experimental Setup

In the previous work [52, 53], we have analyzed variability in the features using three databases - HTIMIT, OGI Stories and TIMIT. In this work, we present results of MANOVA using OGI Stories database; mainly for the comparison with Yang's results [30, 23]. English part of OGI Stories database consists of 207 speakers, speaking for approximately 1 minute each. Each utterance is transcribed at phone level. Therefore, phone is considered as a source of variability or source of information. The utterances are not labeled separately by speakers and channels, so we cannot measure speaker and channel as separate sources. Instead, we assume that different speakers have used different channels and consider speaker+channel as a single source of variability or a single source of information. Features are described in Chapter 2.

G.2 MANOVA

Multivariate analysis of variance (MANOVA) [47] is used to measure the variation in the data, $\{X \in R^n\}$, with respect to two or more factors. In this work, we use two factors - phone and speaker+channel. The underline model of MANOVA is

$$X_{ijk} = \bar{X}_{...} + \bar{X}_{i..} + \bar{X}_{ij.} + \epsilon_{ijk}$$

where, $i = 1, \dots, p$, represents phones, $j = 1, \dots, sc$, represents speakers and channels. This equation shows that any feature vector, X_{ijk} , can be approximated using a sum of $\bar{X}_{...}$, the mean of the data; $\bar{X}_{i..}$, mean of the phone i ; $\bar{X}_{ij.}$, mean of the speaker and channel j , and phone i ; and ϵ_{ijk} , an error in this approximation. Using this model, the total covariance can be decomposed as follows

$$\Sigma_{total} = \Sigma_p + \Sigma_{sc} + \Sigma_{residual} \quad (G.1)$$

where

$$\begin{aligned}\Sigma_p &= \sum_i \frac{N_i}{N} (\bar{X}_{i.} - \bar{X}_{..})^t (\bar{X}_{i.} - \bar{X}_{..}) \\ \Sigma_{sc} &= \sum_i \sum_j \frac{N_{ij}}{N} (\bar{X}_{ij} - \bar{X}_{i.})^t (\bar{X}_{ij} - \bar{X}_{i.}) \\ \Sigma_{residual} &= \frac{1}{N} \sum_i \sum_j \sum_k (X_{ijk} - \bar{X}_{ij})^t (X_{ijk} - \bar{X}_{ij})\end{aligned}$$

and, N is the data size and N_{ijk} refers to the number of samples associated with the particular combination of factors (indicated by the subscript). The computation of the covariance terms is described in Chapter 2

Results of MANOVA are interpreted at two levels - feature element and feature vector. Results for each feature element are shown in Figure G.1. Table G.1 shows the results using the complete feature vector. The contribution of different sources is calculated as $trace(\Sigma_{source})/trace(\Sigma_{total})$. Note that this measure cannot be used to compare variabilities across feature-sets with different number of features. Therefore, we cannot directly compare contribution of variabilities in time and frequency domains. For comparison, contribution of sources in temporal domain is calculated as $trace(E^t \Sigma_{source} E)/trace(E^t \Sigma_{total} E)$, where $E_{101 \times 15}$ is a matrix of 15 leading eigenvectors of Σ_{total} .

In spectral domain, the highest phone variability is between 4-6 Barks. The highest speaker and channel variability is between 1-2 Barks where phone variability is the lowest. In temporal domain, phone variability spreads for approximately 250 ms around the current phone. Speaker and channel variability is almost constant except around the current frame. This deviation is explained by the difference in the phonetic context among the phone instances across different speakers. Thus, features for speakers within a phone differ not only because of different speaker characteristics but also different phonetic contexts. This deviation is also seen in the speaker and channel information in the proposed analysis. In the overall results for each domain, spectral domain has higher variability due to different phones than temporal domain. It also has higher speaker and channel variability than temporal domain.

The disadvantage of this analysis is that it is difficult to interpret the results. For example, how much phone variability is needed for perfect phone recognition? and is 4%

Table G.1: Contribution of sources in spectral and temporal domains

| source | % contribution | |
|-----------------|-----------------|-----------------|
| | Spectral Domain | Temporal Domain |
| phone | 35.3 | 4.0 |
| speaker+channel | 41.1 | 30.3 |

of phone variability in temporal domain significant? In order to answer these questions, we propose an information theoretic analysis.

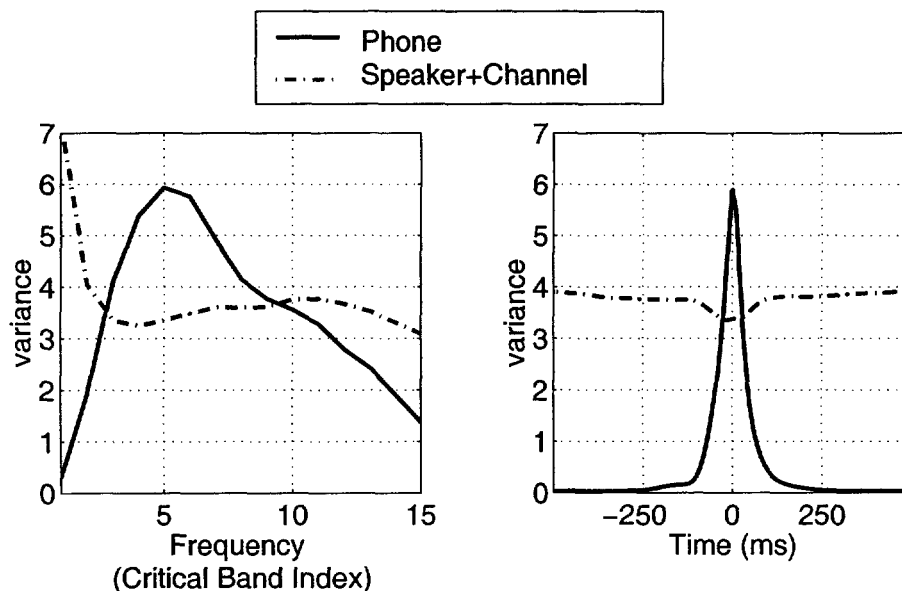


Figure G.1: Results of analysis of variability

G.3 Information-theoretic Analysis

Results of MANOVA can not be directly converted to MI because the determinant of source and residual covariances do not add to the determinant of total covariance. Therefore, we propose a different formulation for the information theoretic analysis as follows. Let $\{X \in R^n\}$ be a set of feature vectors, with probability distribution $p(X)$. Let $h(X)$ be the entropy of X . Let $Y = \{Y_1, \dots, Y_m\}$ be a set of different factors and each Y_i be a set of classes within each factor. For example, we can assume that $Y_1 = \{y_1^i\}$ represents phone factor and each y_1^i represent a phone class. Lets assume that X has two parts; one

completely characterized by Y and another part, Z , characterized by $N(X) \sim \mathcal{N}(0, I_{n \times n})$, where I is the identity matrix. Let $I(X; Y)$ be the MI between X and Y . Assuming that we consider all the possible factors for our analysis,

$$I(X; Y) = I(X; Y_1, \dots, Y_m) = h(X) - h(X/Y_1, \dots, Y_m) = h(X) - h(Z) = D(P||N),$$

where $D(\cdot)$ is the kullback-liebler distance [58] between distributions P and N . Using the chain-rule, the left hand side can be expanded as follows,

$$I(X; Y_1, \dots, Y_n) = I(X; Y_1) + I(X; Y_2/Y_1) + \sum_{i=3}^m I(X; Y_i/Y_{i-1}, \dots, Y_2, Y_1). \quad (\text{G.2})$$

If we assume that there are only two factors Y_1 and Y_2 used for the analysis, then this equation is similar to the decomposition performed using MANOVA (Equation G.1). The term on the left hand side is entropy of X which is the total information in X that can be explained using Y . This is similar to the left-hand side term in MANOVA that describes the total variability. On the right hand side, first term is similar to the phone variability, second term is similar to the speaker variability, and the last term which calculates the effect of unaccounted factors (Y_3, \dots, Y_m) is similar to the residual variability.

First and second terms on the right hand side of Equation G.2 are computed as follows.

$$I(X; Y_1) = h(X) - h(X/Y_1) \quad (\text{G.3})$$

$$I(X; Y_2/Y_1) = h(X/Y_1) - h(X/Y_1, Y_2). \quad (\text{G.4})$$

$h(\cdot)$ terms are estimated using parametric approximation to the total and conditional distribution. It is assumed that the total distribution of features is a Gaussian distribution with covariance Σ . Therefore, $h(X) = \frac{1}{2} \log(2\pi e)^n |\Sigma|$. Similarly, we assume that the distribution of features of different phones (i) is a Gaussian distribution with covariances Σ_i . Therefore,

$$h(X/Y_1) = \frac{1}{2} \sum_{y_1^i \subset Y_1} p(y_1^i) \log(2\pi e)^n |\Sigma_i| \quad (\text{G.5})$$

Finally, we assume that the distribution of features of different phones spoken by different speakers is also a Gaussian distribution with covariances Σ_{ij} . Therefore,

$$h(X/Y_1, Y_2) = \frac{1}{2} \sum_{y_1^i \subset Y_1, y_2^j \subset Y_2} p(y_1^i, y_2^j) \log(2\pi e)^n |\Sigma_{ij}| \quad (\text{G.6})$$

Table G.2: Mutual information between features and phone and speaker and channel labels in spectral and temporal domains

| source | MI (nats) | |
|-----------------|-----------------|-----------------|
| | Spectral Domain | Temporal Domain |
| phone | 1.6 | 1.2 |
| speaker+channel | 0.6 | 5.9 |

Substituting equations G.5 and G.6 in equations G.3 and G.4, we get

$$I(X; Y_1) = \frac{1}{2} \log \frac{|\Sigma|}{\prod_{y_i^i \in Y_1} |\Sigma_i|^{p(y_i^i)}}$$

$$I(X; Y_2/Y_1) = \frac{1}{2} \log \frac{\prod_{y_i^i \in Y_1} |\Sigma_i|^{p(y_i^i)}}{\prod_{y_i^i \in Y_1, y_j^j \in Y_2} |\Sigma_i|^{p(y_i^i, y_j^j)}}$$

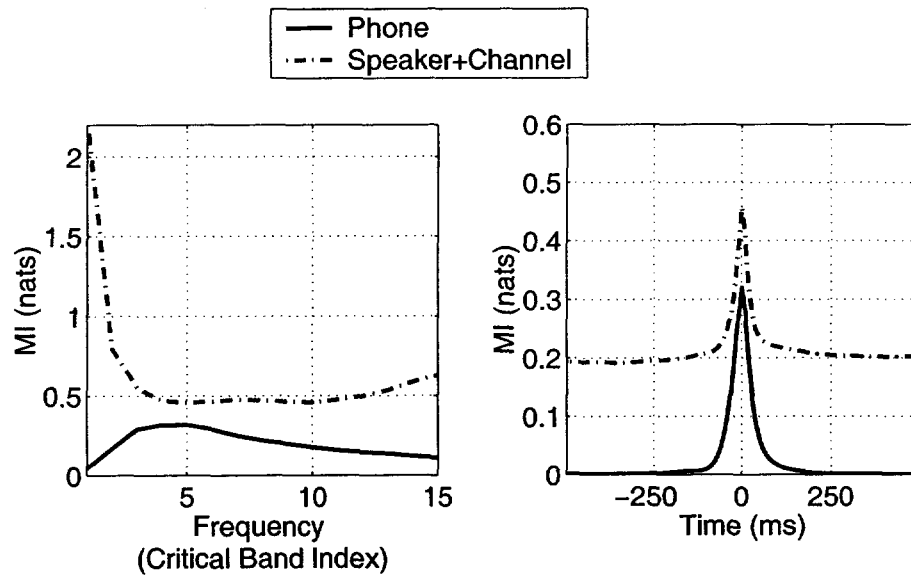


Figure G.2: Results of information-theoretic analysis

Figure G.2 shows the results of information-theoretic analysis in spectral and temporal domain. These results are computed independently for each feature element. In spectral domain, phone information is highest between 3-6 Barks. Speaker and channel information is lowest in that range and highest between 1-2 Barks. Since OGI Stories database was collected over different telephones, speaker+channel information below 2 Barks (≈ 200 Hz) is due to different telephone channels. In temporal domain, the highest phone information

is at the center (0 ms). It spreads for approximately 200 ms around the center. Speaker and channel information is almost constant across time except near the center.

Note that the nature of speaker and channel variability also deviates from the constant around the current frame. But, at the current frame, phone variability is higher than speaker and channel variability. The results of analysis of information show that, at the current frame, phone information is lower than speaker and channel information. This difference is explained by comparing our MI results with results from Yang et. al. [23] in the next section.

Table G.2 shows the results for the complete feature vector. Note that there are some practical issues in computing determinants in Equation G.3 and G.4. They are related to data insufficiency, specifically, in temporal domain where the feature vector is 101 points and there are approximately 60 vectors per speaker per phone. We observe that without proper conditioning of covariances, the analysis overestimates MI ($I(X; Y_1, Y_2) > H(Y_1, Y_2)$). This is addressed using the condition number to limit the number of eigenvalues used in the calculation of determinants. Our hypothesis is that in presence of insufficient data, only few leading eigen vectors are properly estimated. We have use condition number of 1000 to estimate determinant of Σ and Σ_i , and condition number of 100 to estimate the determinant of Σ_{ij} . The results show that phone information in spectral domain is 1.6 nats. Speaker and channel information is 0.5 nats. In temporal domain, phone information is about 1.2 nats. Speaker and channel information is 5.9 nats. Comparison of results from spectral and temporal domains shows that spectral domain has higher phone information than temporal domain. Temporal domain has higher speaker and channel information than spectral domain.

Using these results, we can answer the questions raised in Section G.2. First question was how much phone variability is needed for perfect phone recognition? The answer to the question is $H(Y_1)$, because the maximum value of $I(X; Y_1)$ is $H(Y_1)$. We compute $H(Y_1)$ using phone priors. For this database, we get $H(Y_1) = 3.42$ nats, that means we need 3.42 nats of information for perfect phone recognition. Question about significance of phone information in temporal domain is addressed by comparing it with information-less MI level. The information-less MI is computed as MI between the current phone label

and features at 500 ms in the past or in the future. From our results, we get information-less MI equal to 0.0013 nats considering feature at 500 ms in the past, and 0.0010 nats considering features at 500 ms in the future¹. The phone information in temporal domain is 1.2 bits that is greater than both the levels. Therefore it is significant.

G.4 Results in Perspective

In the proposed analysis, we estimated MI assuming Gaussian distribution for the features. This assumption is validated by comparing our results with the results from a study by Yang, et. al.,[23], where MI was computed without assuming any parametric model for the distribution of features. Note that only entropies can be directly compared for difference in the estimation technique [58]. However, MI using Gaussian assumption can be equal to, less or more than the actual MI. In the comparison of our results with Yang's results, we consider only the nature of information observed in both studies. The difference in actual MI levels across the two studies is related to the difference in the estimation techniques.

In spectral domain, Yang's study showed higher phone information between 3-8 Barks. The highest phone information was observed at 4 Barks. Higher speaker and channel information was observed around 1-2 Barks. In temporal domain, their study showed that phone information spreads for approximately 200 ms around the current time frame. Comparison of results from this analysis and our analysis shows that nature of phone information is similar in both studies. Nature of speaker and channel information in spectral domain is also similar. We could not compare the speaker and channel information in temporal domain because Yang's study did not present these results.

In Section G.3, we observed difference in the nature of speaker and channel variability, and speaker and channel information at $f_i = 5$ Barks. Comparing MI levels from our study to those from Yang's study, we observe that Yang's results show that speaker and channel information at 5 Barks is less than the corresponding phone information. This is consistent with results of analysis of variability, but not with the proposed analysis of information. As mentioned before, this difference is due to difference in the density

¹Information-less MI calculated using Yang et. al. is 0.019 bits

estimation techniques used for computing MI. In the future work, we plan to model the densities using more sophisticated techniques, and improve the estimation of speaker and channel information.

G.5 Conclusions

We proposed analysis of information in speech using three sources of information - language (phone), speaker and channel. Information in speech was measured as MI between the class labels and the set of features extracted from speech signal. For example, linguistic information was measured using phone labels and the features. We modeled distribution of features using Gaussian distribution. Thus we related the analysis to previous proposed analysis of variability in speech. We observed similar results for phone variability and phone information. The speaker and channel variability and speaker and channel information around the current frame was different. This was shown to be related to the over-estimation of speaker and channel information using unimodal Gaussian model. Note that the analysis of information was proposed because its results have more meaningful interpretations than results of analysis of variability. For addressing the over-estimation, we plan to use more complex models, such as mixture of Gaussians, for computing MI in the future work.

Bibliography

- [1] E. S. PARRIS AND M. J. CAREY. Discriminative Phonemes for Speaker Identification . In *Proc. of ICSLP* (Yokohama, Japan, 1994), pp. 1843–1846.
- [2] J. A. BILMES. Maximum Mutual Information Based Reduction Strategies for Cross-correlation Based Joint Distribution Modelling . In *Proc. of ICASSP* (Seattle, USA, 1998), pp. 469–472.
- [3] J. P. EATOCK AND J. S. MASON. A Quantitative Assessment of the Relative Speaker Discriminating Properties of Phonemes. In *Proc. of ICASSP* (Adelaide, Australia, 1994), pp. I133–136.
- [4] K. L. BROWN AND E. B. GEORGE. CTIMIT: A speech corpus for the cellular environment with applications to automatic speech recognition . In *Proc. of ICASSP* (Michigan, USA, 1995), pp. 105–108.
- [5] A. ADAMI, L. BURGET, S. DUPONT, H. GARUDADRI, F. GREZL, H. HERMANSKY, P. JAIN, S. KAJAREKAR, N. MORGAN AND S. SIVADAS. QUALCOMM-ICSI-OGI Features for ASR. In *To appear in Proc. of ICSLP* (Colorado, USA, 2002).
- [6] A. ADAMI, S. KAJAREKAR AND H. HERMANSKY. A new speaker change detection method for two-speaker segmentation. In *Proc. of ICASSP* (Orlando, USA, 2002), pp. 3908–3910.
- [7] A. MORRIS, J. L. SCHWARTZ, AND P. ESCUDIER. An information theoretical investigation into the distribution of phonetic information across the auditory spectrogram. *Computer Speech and Language* 7 (April 1993), 121–136.
- [8] C. AVENDANO, S. VAN VUUREN AND H. HERMANSKY. Data-Based RASTA-Like Filter Design for Channel Normalization in ASR. In *Proc. of ICSLP* (Philadelphia, USA, 1996), pp. 2087–2090.
- [9] C. AVENDANO, S. VAN VUUREN AND H. HERMANSKY. Optimizing RASTA filters on corrupted speech. In *Proc. of ICSLP* (Philadelphia, USA, 1996), pp. 889–892.
- [10] D. C. HOWELL. *Statistical methods in Psychology*. Duxbury Press, Boston, 1989.

- [11] D. O'SHAUGHNESSY. Speaker recognition. *IEEE ASSP Magazine* 3 (1986), 4–17.
- [12] D. REYNOLDS. Speaker identification and verification using gaussian mixture models. *Speech Communication* 17 (1995), 91–108.
- [13] D. REYNOLDS. HTIMIT and LLHDB : Speech Corpora for the Study of Handset Transducer Effects. In *Proc. of ICASSP* (Munich, Germany, 1997), pp. 1535–1538.
- [14] D. REYNOLDS, T. QUATIERI AND R. DUNN. Speaker verification using adapted mixture models. *Digital Signal Processing* 10 (2000), 181–202.
- [15] D. X. SUN AND L. DENG. Analysis of acoustic-phonetic variations in fluent speech using timit. In *Proc. of ICASSP* (Detroit, USA, 1995), pp. 201–204.
- [16] E. ZWICKER AND E. TERHARDT. Analytical expressions for critical band rate and critical bandwidth as a function of frequency. *The Journal of The Acoustical Society of America* 68 (1980), 1523–1525.
- [17] F. ITAKURA. Line spectrum representation of linear predictive coefficients of speech signal. *The Journal of The Acoustical Society of America* 57 (1975), S35.
- [18] F. JELINEK. *Statistical Methods for Speech Recognition*. MIT Press, Cambridge, 1997.
- [19] F. WEBER, B. PESKIN, M. NEWMAN, A. CORRADA-EMMANUEL AND L. GILLICK. Speaker recognition on single- and multispeaker data. *Digital Signal Processing* 10 (2000), 75–92.
- [20] G. E. P. BOX AND S. L. ANDERSON. Permutation theory in the development of robust criteria and the study of departures from assumptions. *Journal of the Royal Statistical Society Series B* 17 (1955), 1–34.
- [21] G.R. DODDINGTON. Speaker recognition—identifying people by their voices. *Proceedings of the IEEE* 11 (1985), 1651–1664.
- [22] H. BOURLARD AND N. MORGAN. *Connectionist Speech Recognition – A Hybrid Approach*. Kluwer Academic Publishers, Boston, 1994.
- [23] H. H. YANG, S. SHARMA, S. VAN VUUREN AND H. HERMANSKY. Relevance of Time-Frequency Features for Phonetic and Speaker-Channel Classification. *Speech Communication* 31 (May 2000), 35–50.
- [24] H. HERMANSKY. Perceptual linear predictive (PLP) analysis for speech. *The Journal of The Acoustical Society of America* 87 (Apr. 1990), 1738–1752.

- [25] H. HERMANSKY. Should recognizers have ears? *Speech Communication* 25, 1-3 (1998), 3-27.
- [26] H. HERMANSKY AND N. MALAYATH. Spectral basis functions from discriminant analysis. In *Proc. of ICSLP* (Sydney, Australia, 1998), pp. 1379-1382.
- [27] H. HERMANSKY AND N. MORGAN. RASTA processing of speech. *IEEE Trans. Speech and Audio* 2 (Oct. 1984), 578-589.
- [28] H. HERMANSKY, N. MORGAN, A. BAYYA AND P. KOHN. Rasta-plp speech analysis technique. In *Proc. of ICASSP* (San Francisco, USA, 1992), pp. 121-124.
- [29] H. R. LINDMAN. *Analysis of variance in experimental design*. Springer-Verlag, New York, 1992.
- [30] H. YANG, S. VAN VUUREN AND H. HERMANSKY. Relevancy of Time-Frequency Features for Phonetic Classification Measured by Mutual Information. In *Proc. of ICASSP* (Phoenix, USA, 1999), pp. 225-228.
- [31] J. KOOLWAAIJ AND J. DE VETH. The use of broad phonetic class models in speaker recognition. In *Proc. of ICSLP* (Sydney, Australia, 1998), pp. 3357-3362.
- [32] J. PELECANOS AND S. SRIDHARAN. Feature warping for robust speaker verification. In *Proc. of 2001: A Speaker Odyssey : The Speaker Recognition Workshop* (Crete, Greece, 2001), pp. 213-218.
- [33] K. FUKUNAGA. *Statistical Pattern Recognition*. Academic Press, Indiana, 1990.
- [34] L. R. BAHL, P. V. DE SOUZA, P. S. GOPALAKRISHNAN, D. NAHAMOO, AND M. A. PICHENY. Robust methods for using context-dependent features and models in a continuous speech recognizer. In *Proc. of ICASSP* (Adelaide, Australia, 1994), pp. 533-536.
- [35] L. R. BAHL, S. BALAKRISHNAN-AIYER, J. BELLEGARDA, M. FRANZ, P. GOPALAKRISHNAN, D. NAHAMOO, M. NOVAK, M. PADMANABHAN, M. PICHENY AND S. ROUKOS. Performance of the IBM large vocabulary continuous speech recognition system on the ARPA wall street journal task. In *Proc. of ICASSP* (Detroit, USA, 1995), pp. 41-44.
- [36] L. RODRIGUEZ-LUNARES AND C. GARCIA-MATEO. On the use of acoustic segmentation in speaker identification. In *Proc. of EUROSPEECH* (Rhodos, Greece, 1997), pp. 2315-2318.

- [37] L.R. RABINER. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE 77* (Feb. 1989), 257-285.
- [38] M. HUNT. Spectral Signal Processing for ASR. In *Proc. of ASRU* (Colorado, USA, 1999), pp. 17-25.
- [39] M. HUNT AND C. LEFEBVRE. A comparison of several acoustic representations for speech recognition with degraded and undegraded speech. In *Proc. of ICASSP* (Glasgow, Scotland, 1989), pp. 262-265.
- [40] M. HUNT, S. RICHARDSON, D. BATEMAN AND A. PIAU. An investigation of plp and imelda acoustic representations and of their potential for combination. In *Proc. of ICASSP'91* (Toronto, Canada, 1991), pp. 881-884.
- [41] M.J. HUNT. A statistical approach to metrics for word and syllable recognition. *The Journal of The Acoustical Society of America 66* (1979), S1-S35.
- [42] N. MALAYATH. *Data-Driven Methods for Extracting Features from Speech*. PhD thesis, Oregon Graduate Institute of Science and Technology, Portland, USA, Jan. 2000.
- [43] N. MALAYATH, H. HERMANSKY, S. KAJAREKAR AND B. YEGNANARAYANA. Data-driven temporal filters and alternatives to GMM in speaker verification. *Digital Signal Processing 10* (2000), 55-74.
- [44] P. MCMAHON, N. HARTE, S. VASEGHI AND P. MCCOURT. Discriminative spectral-temporal multiresolution features for speech recognition. In *Proc. of ICSLP* (Phoenix, USA, 1999), pp. 1055-1058.
- [45] R. A. COLE AND J. MARIANI AND H. USZKORIET AND A. ZAENEN AND V. ZUE, Ed. *Survey of the State of the Art in Human Language Technology*. Cambridge University Press, Standford University, Stanford, CA, 1996.
- [46] R. COLE, M. NOEL AND T. LANDER. Telephone speech corpus development at CSLU. In *Proc. of ICSLP* (Yokohama, Japan, 1994), pp. 821-824.
- [47] R. V. HOGG AND E. A. TANNIS. *Statistical Analysis and Inference*, fifth ed. PRANTICE HALL, New Jersey, 1997.
- [48] S. FURUI. On the role of spectral transition for speech perception. *The Journal of The Acoustical Society of America 80* (Apr. 1986), 1016-1025.

- [49] S. FURUI. Speaker-independent isolated word recognition based on emphasized spectral dynamics. In *Proc. of ICASSP* (Tokyo, Japan, 1986), pp. 1991–1994.
- [50] S. K. GUPTA AND MICHAEL SAVIC. Text-independent speaker verification based on broad phonetic segmentation of speech. *Digital Signal Processing 2* (1992), 181–202.
- [51] S. KAJAREKAR AND H. HERMANSKY. Analysis of information in speech and its application in speech recognition. In *Proc. of TSD* (Brno, Czech Republic, 2000), pp. 283–288.
- [52] S. KAJAREKAR, N. MALAYATH AND H. HERMANSKY. Analysis of sources of variability in speech. In *Proc. of EUROSPEECH* (Budapest, Hungary, 1999), pp. 343–346.
- [53] S. KAJAREKAR, N. MALAYATH AND H. HERMANSKY. Analysis of speaker and channel variability in speech. In *Proc. of ASRU* (Colorado, USA, 1999), pp. 59–62.
- [54] S. S. STEVENS AND J. VOLKMANN. The relation of pitch of frequency: A revised scale. *American Journal of Psychology 53* (1940), 329–353.
- [55] S. SHARMA, D. ELLIS, S. KAJAREKAR, P. JAIN AND H. HERMANSKY. Feature extraction using non-linear transformation for robust speech recognition on the AU-RORA data-base. In *Proc. of ICASSP* (Istanbul, Turkey, 2000), pp. 1117–1120.
- [56] S. VAN VUUREN AND H. HERMANSKY. Data-driven design of rasta-like filters. In *Proc. of EUROSPEECH* (Rhodes, Greece, 1997), pp. 409–412.
- [57] S.B. DAVIS AND P. MERMELSTEIN. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustic, Speech and Signal Processing 28* (Aug. 1980), 357–366.
- [58] T. M. COVER AND J. A. THOMAS. *Elements of Information theory*. John Wiley & Sons, Inc., New York, 1991.
- [59] T. NITTA. A novel feature-extraction for speech recognition based on multiple acoustic-feature planes. In *Proc. of ICASSP* (Seattle, USA, 1998), pp. 29–32.

Biographical Note

Sachin S. Kajarekar was born in Mumbai, Maharashtra, India, on the 11th of August 1973. He completed Bachelor of Electronics Engineering degree in 1994 from Father Agnel College of Engineering, Bombay University. In his first job as a software engineer, he worked on feature enhancement and bug fixing of Sequoia fault-tolerant Unix operating system. He was one of the first employees of Informix India Development Center where his job included technology transfer from Informix Singapore Development Center. From 1996 to 1997, he worked as a software engineer at Oracle Corp., USA, where he worked in Sequent-Oracle product development group.

In September 1997 he joined Oregon Graduate Institute as M.S. Student. During M.S., he worked with different professors in areas such as speech recognition, speech synthesis, and adaptive signal processing. In the summer of 1998, he joined Anthropic Signal Processing Group as Ph.D. student. Under the guidance of Prof. Hermansky, he worked on several research projects which include, phone-based speaker verification system and design of two-dimensional linear discriminants for robust speech recognition. During his Ph.D, he has authored and co-authored several conference papers.