Data-Driven Methods for Extracting Features from Speech

Narendranath Malayath
M.S., Indian Institute of Technology, Madras, India, 1995

A dissertation submitted to the faculty of the
Oregon Graduate Institute of Science and Technology
in partial fulfillment of the
requirements for the degree
Doctor of Philosophy
in
Department of Electrical and Computer Engineering

January 2000

The dissertation "Data-Driven Methods for Extracting Features from Speech" by Narendranath Malayath has been examined and approved by the following Examination Committee:

Dr. Hynek Hermansky
Professor
Thesis Research Adviser

Dr. Harinath Garudadri
Staff Engineer
Qualcomm Inc.

Dr. Michael Macon
Assistant Professor

Dr. Lorin Netsch
Senior Member of the Technical Staff
Texas Instruments, Inc.

Dr. Howard Yang
Assistant Professor

Dedication

To The Science of Speech Communication

Acknowledgements

I express my at most gratitude to Prof. Hynek Hermansky, my research advisor, for the numerous technical discussion that I had with him. Many of the ideas presented in this dissertations are the result of these discussions. I have learned from Hynek, the true spirit of scientific research. I also thank Hynek for supporting me through research grants throughout my stay at OGI.

I thank Dr. Harinath Garudadri, Dr. Michael Macon, Dr. Lorin Netsch and Dr. Howard Yang for serving in my thesis committee. The revisions suggested by them have improved the quality of the dissertation.

During my three year long stay at OGI, the members of the Anthropic signal processing group have contributed to this work through helpful suggestions and comments. Special thanks go to Sachin for the discussions that we had on almost a daily basis. These discussions have helped me in clearly formulating many of the problems addressed in this dissertation. I also thank Sunil for his suggestions which improved the readability of the dissertation.

I acknowledge the help provided by the Center for Spoken Language Understanding (CSLU) in the form of software tools and speech databases. I thank Sarel for the software which was used for all the speaker verification experiments reported in this dissertation. I also acknowledge the help provided by Pavel Chytil by promptly maintaining the computational resources of the Anthropic signal processing group.

Contents

List of Tables

List of Figures

Abstract

Data-Driven Methods for Extracting Features from Speech
Narendranath Malayath, *B.Tech.*, M.S.

Ph.D., Oregon Graduate Institute of Science and Technology
January, 2000

Thesis Research Advisor: Dr. Hynek Hermansky

Feature extraction plays a major role in any form of pattern recognition. Current feature extraction methods used for automatic speech recognition (ASR) and speaker verification rely mainly on properties of speech production (modeled by all-pole filters) and perception (critical band integration simulated by Mel/Bark filter bank). We propose stochastic methods to design feature extraction methods which are trained to alleviate the unwanted variability present in speech signal. In this dissertation we show that such data-driven methods provide significant advantages over the conventional methods for feature extraction.

In the first part of the dissertation discriminant methods are introduced for extracting spectral features for ASR. Spectral basis functions which preserve phonetic class separability are derived using linear discriminant analysis (LDA). It is observed that the discriminant basis functions analyze the low frequency part of the spectrum with higher resolution than the high frequency part. This trend is consistent with properties of human hearing which are explained using the notion of critical bandwidth and emulated in the current feature extraction modules by Mel/Bark filter bank. The proposed discriminant

features are shown to outperform the conventional features in ASR experiments.

The second part of the dissertation introduces data-driven methods for the design of channel normalizing filters for speaker verification. It has been observed that a reasonable verification error can be achieved if the speaker uses the same handset and telephone line for testing. On the other hand if the speaker uses a different telephone handset while testing, the verification error can increase by four to five times. We introduce a data-driven method for designing filters capable of normalizing the variability introduced by different telephone handsets. The design of the filter is based on the estimated second order statistics of handset variability. This filter is applied on the logarithmic energy outputs of Mel spaced filter banks. The effectiveness of the proposed channel normalizing filter in improving speaker verification performance in mismatched conditions is also demonstrated.

Chapter 1

Introduction

Automatic extraction of useful information from speech has been a subject of active research for many decades. It is known that the information contained in speech is carried from the speaker to the listener in the form of vibrations (pressure waveform) in air. These vibrations can be converted into an electric signal by a microphone which acts as a transducer. Typically, it is the sampled and digitized form of this electric signal that we refer to as the *speech signal*. Automatic speech recognition(ASR) focuses on the extraction of the linguistic message from the signal. ASR could lead to the development of more efficient and natural means of communication with computers. It could also play a crucial role in making business over the telephone more efficient and cost effective [2]. Speaker verification addresses the problem of verifying the identity of a speaker from the speech signal. Speaker verification can be used for verifying the identity of a customer in banking applications or automatic tagging of speech in multi-speaker conversations [3, 4, 5, 6, 7].

The speech signal contains information about many sources. Information about these sources get manifested in the signal as variability. For example, the signal characteristics vary when a speaker utters two different phonemes. Knowledge of this variability and its relation with the spoken phonemes enable us to predict the phoneme from the speech signal. The relation between the source of information and the variability that it causes can be represented by class-conditional distributions in the signal space. The farther apart the class-conditional distributions are located in the signal space the better would be the prediction. This prediction cannot be error free if there are other sources of information causing variability which make the class-conditional distributions overlap.

The application dictates the source of information that needs to be focused on. For

example, in ASR it is the *linguistic information* that is of interest while the variability due to speaker information can cause confusion and needs to be treated as noise. Similarly, for speaker verification speaker information is important while variability due to microphones information could potentially degrade the performance. Hence it is evident that recognition systems need a mechanism to focus on the desired source of information while ignoring undesired sources of information.

One method to achieve this is by training stochastic classifiers with data that contain the various sources of information. For example speaker independence is achieved in ASR by training the recognizer using speech data collected from multiple speakers. This approach has the following disadvantages:

- A large amount of data is necessary to achieve a good coverage of all the sources of information. Additionally, training the classifier with large amount of data is computationally expensive.

- If a particular source of variability, for example, acoustic environment (or speaker) is not well covered in the training data then a new acoustic environment (or speaker) makes it necessary to retrain the recognizer. In addition to retraining, this would also necessitate fresh data collection in the new acoustic environment or from the new speaker.

Incremental adaptation of the stochastic classifiers using a limited amount of data is an effective alternative to retraining the classifier. Adaptation methods are often used to adapt parameters of the classifier to the voice of a new user or a microphone [8, 9]. Yet another method involves extraction of features from the speech signal that carry primarily information about the desired source while suppressing the information about the undesired sources [10]. Training the stochastic classifiers, adaptation to new acoustic environments or speakers, and the use of robust features are all important in providing robustness in recognition systems [1].

The feature extraction methods that are currently used in speech and speaker recognition systems are often based on human knowledge about speech production and perception.

[1] We use the term robustness to mean invariance to undesired sources of information.

For example, linear predictive coding (LPC) of speech is based on modeling the speech production mechanism as an all-pole filter [11]. Similarly, Mel frequency cepstral analysis [12, 13] and perceptual linear predictive analysis [14] are motivated by properties of human auditory system. Incorporating properties of speech production and auditory perception could improve speech processing. However, the specific properties of speech production and auditory perception which are important for reliable recognition of speech and speaker are not clearly understood. In other words, all the properties of speech production and perception may not be relevant (or may be even harmful) for extracting information from speech. Moreover, the objective of feature extraction is not to model the speech production accurately or to mimic human auditory perception, but we are interested in features which carry primarily information about the desired source of information while suppressing the information about the undesired sources.

In this thesis, we propose novel feature extraction methods that are designed to focus on the desired source of information while de-emphasizing the undesired sources of information. These feature extraction methods are trained on databases that are independent of any specific tasks and hence need not be retrained for every new task. We demonstrate that the use of such data-driven feature extraction methods improves the performance of speech and speaker recognition systems while providing useful knowledge about the nature of information carried by the speech signal.

The remaining part of the chapter is organized as follows. The various sources of information that are carried by the speech signal are discussed in the next section. The relation between the variability introduced by a source and the information contained in the signal about the source is established in Section 1.2. Relevance of extracting features from the signal is emphasized in Section 1.3. Section 1.4 gives a brief introduction to the current methods for speech and speaker recognition. Section 1.5 introduces the notion of data-driven feature extraction. The issues that need to be addressed in data-driven feature extraction are discussed in Section 1.5.1. Section 1.6 summarizes the contributions of the dissertation. Section 1.7 gives an overview of the dissertation.

1.1 Sources of Information in Speech

The characteristics of the speech signal (or measurements made from the signal) vary with many factors. For example, if the same speech sound uttered by the same person is recorded using different microphones, the resulting set of signals will have different characteristics depending upon the nature of the microphones. Hence the probability distribution of the signal depends on the microphone used and the signal contains information about the microphone. The main sources of information in speech signal are the following: (a) Phoneme, (b) Phonetic context or Coarticulation (c) Speaker, (d) Channel.

Phoneme Information

The smallest segments of sound that can cause a change in the meaning of a word are termed as *phonemes* [15]. Often phonemes are considered as the basic units in conveying linguistic information and hence the ability to correctly hypothesize the phoneme from the acoustic signal is crucial in automatic speech recognition. The information carried by the speech signal about the identity of the phonemes is referred to as *phonetic information.*

Coarticulation-articulation or Context Information

Both the acoustic characteristics and the perception of a sound are influenced by the sounds that precede and follow. The acoustic influence of a sound on its neighbors is mainly caused by the inherent inertia of the speech production mechanism. We call the information carried by a speech sound about the identity of the following and preceding sounds as *context information* or *coarticulation information.*

Speaker Information

Human beings are able to recognize a familiar speaker from speech. The speech signal carries information about many factors that helps us in recognizing speakers. These factors include dialect, prosodic features that are specific to a speaker, vocal tract characteristics and vocal source characteristics [16]. The information that is contained in the signal that carries the identity of a speaker is termed as *speaker information.*

Channel Information

The recording instruments and the communication channel affect the characteristics of the speech signal. The variability introduced by the microphones, communication channels, and any acoustic noise that gets added to the signal is referred to as the *channel information* carried by the speech signal.

1.2 Variance and Information

From the discussion in the previous section it is clear that the speech signal contains information about various sources. In this dissertation we often refer to the variability introduced by a source to imply the presence of information about the source. Often the feature extraction module achieves the goal of suppressing the unwanted sources of information by alleviating the variance introduced by those sources, while preserving the variability introduced by the wanted source. Hence in this section we formally review the relationship of the statistical quantity of variance to the mathematical definition of information. The entropy of a random variable x, $H(x)$, is defined by the expected value of $-\ln[p(x)]$, where $p(x)$ is the probability density function of the random variable x:

$$H(x) = -\int \ln[p(x)]p(x)dx.$$

If the random variable x is Gaussian, then the entropy H is related to the variance parameter σ^2 by the monotonic function [17]

$$H(x) = \ln \sigma\sqrt{2\pi e}.$$

This indicates that the uncertainty associated with normal random variable x monotonically increases with its variance.

The information carried by the speech signal about a source can be quantified by the mutual information between the source and any measurement taken from the speech signal. Mutual information is a measure of the reduction of uncertainty of one random variable due to the knowledge of the other [18]. For example let us consider speaker as the source of information represented by the random variable x. Let the number of speakers be

restricted to two which makes x a discrete random variable that can take two values with equal probability of 0.5. Let y be a single dimensional feature extracted from the speech signal. Then the mutual information $I(x;y)$ tells us about the reduction in uncertainty about the identity of the speaker if we have the knowledge of the measurement y. Hence the mutual information between x and y directly quantifies the amount of information carried by the measurement y about the identity of the speaker. The mutual information $I(x;y)$ is defined by

$$I(x;y) = -\int p(x,y) \ln \left[\frac{p(x,y)}{p(x)p(y)} \right] dxdy. \tag{1.1}$$

If the class conditional probability density functions of the two classes are normal and the the variances of the two distributions are equal then the joint distribution x and y is given by

$$p(x,y) = \sum_{i=1}^{2} 0.5\delta(x-i)N(y:\mu_i,\sigma^2), \tag{1.2}$$

where $N(y:\mu_i,\sigma^2)$ represents a univariate Gaussian function of y parameterized by mean, μ_i and variance σ^2. The categorical variable, x, representing the speaker identity is allowed to take either the value 1 or 2. From equation 1.1 and 1.2 the mutual information between x and y can be shown to be equal to

$$I(x;y) = 0.5 \sum_{i=1}^{2} \int N(y:\mu_i,\sigma^2) \ln \left[\frac{2N(y:\mu_i,\sigma^2)}{\sum_{i=1}^{2} N(y:\mu_i,\sigma^2)} \right]. \tag{1.3}$$

$I(x;y)$ is dependent only on the class conditional means μ_1, μ_2 and the variance σ^2.

Figure 1.1 shows normalized $I(x;y)$ as a function of the sample variance of the class conditional means. It is evident from the figure that for a fixed class conditional variance, an increase in the variance of the class conditional means results in an increase in mutual information. This observation is intuitive since an increase in the separation between the two class conditional means will result in an improved class separability. If the two classes have the same class conditional mean then the measurement y provides no information about the factor x. This follows from the fact that if the variance of the class conditional means is zero the mutual information also diminishes to zero. From the figure it is also

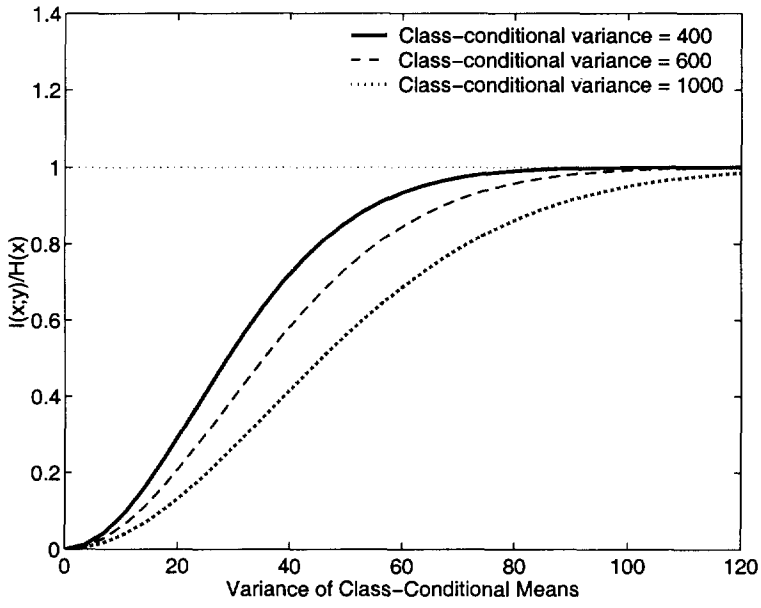


Figure 1.1: Plot showing the effect of class conditional means and variances on the mutual information between the measured quantity y and the classes represented by the categorical variable x.

clear that, if the variance of the class conditional means is held constant, then an increase in the class conditional variance decreases the mutual information.

From the above discussion it is concluded that the variance of the class conditional means is directly related to the information contained in the measurement about the classes. Moreover the class conditional variances can be considered as the noise process that reduces the information contained in the measurement. It is also interesting to note that the variance of the class conditional means (between-class variance) and the average of the class conditional variances (within-class variance) are used in Linear Discriminant Analysis (LDA) to estimate directions in the feature space that preserve maximum class separability. LDA is extensively used in our effort to derive improved features. A detailed discussion on LDA is given in chapter 2.

1.3 Relevance of Feature Extraction

From the discussion presented in the previous section it is clear that there are various sources that introduce variability in the speech signal. Depending upon the application, typically we are interested in identifying only one source of variability. Hence an effective feature extraction technique should be able to capture the variability in the data caused by a desired source while suppressing the variability caused by undesirable sources. For example, in speech recognition, it is highly desirable to have features that carry mainly phonetic information. Similarly for speaker recognition it is important to have features that carry mainly the speaker specific information while alleviating the variability introduced by channel.

Feature extraction cannot provide any significant advantage if, (a) the training data used to train the recognizer contain all the variability(unwanted and wanted) that the recognizer is expected to encounter during the test phase, and (b) the recognizer is flexible enough to capture any arbitrary distribution. In such a case the classifier can directly learn the invariant features from the training data. For the effective functioning of statistical classifiers many constraints are built in. For example, a limited number of Gaussian functions with diagonal covariances are used to approximate class conditional distributions. In many applications it is also not practical to train a classifier using data with all the variability that may be present in the testing phase. For example, in speaker verification, a statistical model has to be built from a relatively small amount of speech data to represent the acoustic features of the target speaker and it is not practical to have training utterances collected from multiple telephone handsets. Hence the model will be highly biased towards the handset used to record the training utterance. Thus in such a situation it is highly desirable to have features that are insensitive to channel characteristics.

1.4 Extracting Information from Speech

Out of the various types of information carried by the speech signal, the phonetic and speaker information are the most important sources. Extraction of phonetic information from the signal is crucial in implementing automatic speech recognition (ASR) systems.

Similarly, extraction of speaker information is inevitable for speaker verification systems. This section briefly discusses the current dominant methods used for ASR and speaker verification.

1.4.1 Automatic Speech Recognition

The goal of automatic speech recognition is to recognize the sequence of symbols from the speech signal. The speech signal is processed first to derive a sequence of feature vectors represented by $X = \mathbf{x}_1\mathbf{x}_2\ldots\mathbf{x}_T$. Then the task is to determine the sequence of symbols spoken, $W = W_1W_2\ldots W_k$. According to Bayes' theory, for minimum error these symbols are to be chosen to maximize the posterior probability

$$P(W|X) = \frac{p(X|W)p(W)}{p(X)}. \tag{1.4}$$

In most current speech recognizers, the class conditional distributions of the feature vectors are modeled using Hidden Markov Models [19, 2]. The prior probability $p(W)$ is implemented as a language model that prefers more probable sequences of symbols over the less probable ones. Hidden Markov Models (HMMs) are trained using sequences of feature vectors extracted from symbols or phonemes. For large vocabulary continuous speech recognition (LVCSR) the language model is trained on a large text corpus.

While both the language model and the HMMs are trained, the feature extraction module is still largely left for the designer to hand-craft.

1.4.2 Speaker Verification

The objective of a speaker verification system is to verify the identity claim of a speaker from his or her speech. A speaker verification system could be either text dependent or it could be text independent [5]. In text dependent speaker verification, the speaker is constrained to speak a specific text. No such restriction is placed in a text independent speaker verification task. Current text-independent speaker verification systems are based on modeling the feature vectors of the target speaker using a Gaussian mixture model (GMM) [20]. The unwanted linguistic variability of the acoustic features is suppressed by the simultaneous use of two GMMs, one is a speaker-specific GMM (λ_s), modeling

the acoustic space of a given speaker, the second is the so called universal background model (λ_b) [21]. The universal background model (UBM) is a speaker independent model, trained using the speech data of a large number of speakers. Thus the UBM represents a speaker independent distribution of the feature vectors. The speaker-specific model is obtained by maximum a posteriori (MAP) adaptation of the UBM with the data of the specific target speaker.

During the verification phase, the claim is rejected or accepted by comparing the log likelihood ratio with a threshold θ as illustrated by the following equation,

$$\ln \frac{p(X|\lambda_s)}{p(X|\lambda_b)} \underset{accept}{\overset{reject}{\lessgtr}} \theta,$$

where X is a set of feature vectors extracted from the speech signal. It has been observed that a reasonable verification error can be achieved if the speaker uses the same handset and telephone line for testing [20]. On the other hand, if the speaker uses a different telephone handset while testing, the verification error can increase by four to five times [22]. Hence processing techniques that are robust to channel variability are of significance in speaker verification.

1.5 Data-driven Feature Extraction - Deriving knowledge from data

Figure 1.2 shows the block diagram of a generic classifier that could be used for ASR or speaker recognition. The speech signal is first processed to extract features and these features are input to a statistical classifier. During the training phase the classifier estimates class conditional probability distributions of the feature vectors. During testing, given a feature vector (or a sequence of vectors) the classifier selects a class (or sequence of classes) that best explains the generation of the feature vectors. The classifier is typically trained on a task specific database. For example, if we are interested in recognizing a sequence of digits, the classifier is trained using a database that consists of continuously spoken digits. If the recognizer has to be speaker independent, then the database needs to have utterances from multiple speakers. Similarly, to make the system robust to handset

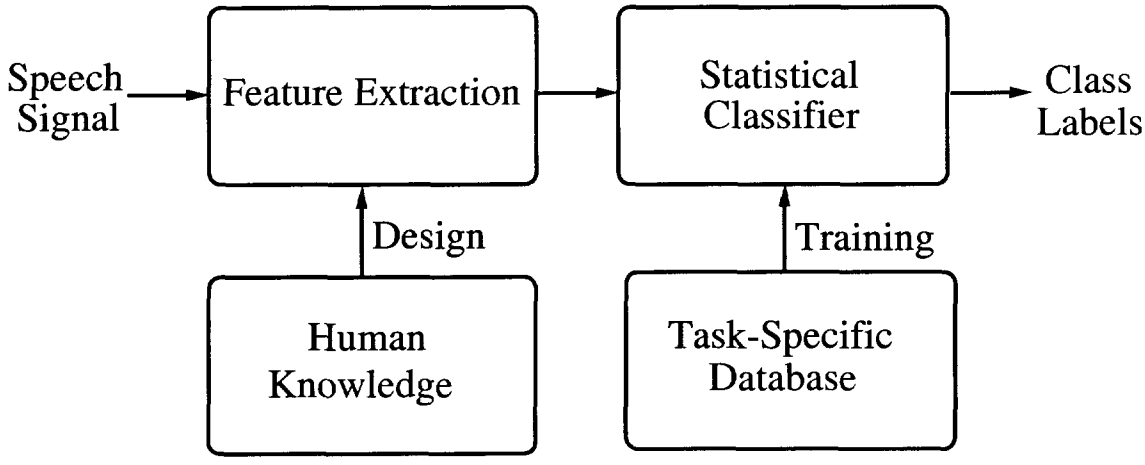


Figure 1.2: Generic form of a recognizer used in speech and speaker recognition. While the statistical classifier is trained using a task specific speech data-base, the design of the feature extraction module is largely left to the knowledge of the designer.

variability, the recognizer has to be trained using utterances recorded through multiple telephone handsets.

While the classifiers are trained on a task-specific database, the feature extraction modules are generally hand crafted to simulate certain properties of human hearing. For example, the short-time Fourier spectrum is passed through a set of auditory-like filters to simulate the non-uniform frequency resolution observed in human auditory system. Even though feature extraction modules based on human speech perception have shown robustness to additive and convolutive distortions [12, 14, 23] they are not necessarily optimized to perform specific tasks like speech recognition and speaker verification.

Figure 1.3 shows the block diagram of the generic recognizer with the proposed data-driven method for feature extraction. The feature extraction module is optimized using a speech corpus that contains multiple sources of variability. The feature extractor is trained to preserve the desired source of variability while suppressing the unwanted variability. The proposed data-driven approach for deriving features can be defined more formally as follows. Let $\mathbf{s}$ be the initial representation of the speech signal. The process of extracting

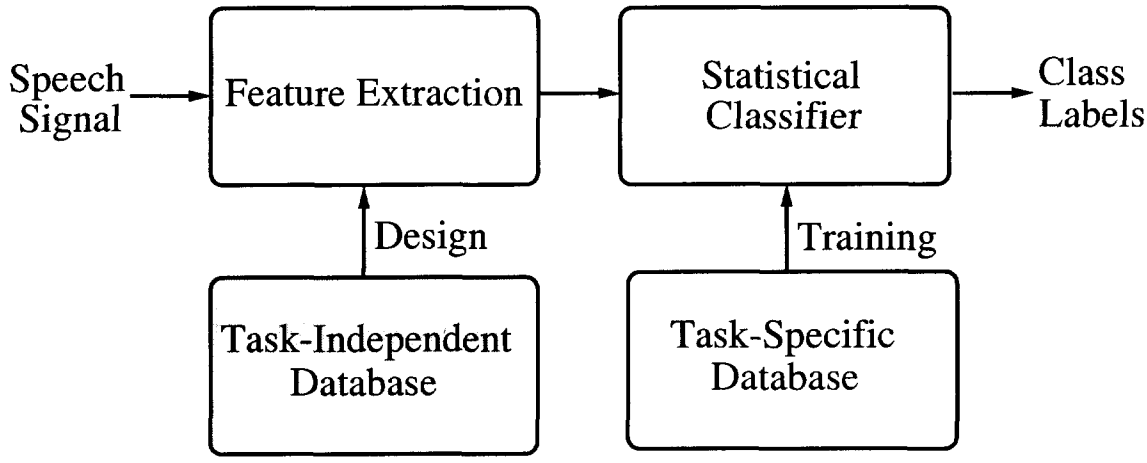


Figure 1.3: The proposed data-driven approach for designing the feature extraction module.

a set of N features, $\mathbf{f}$, from $\mathbf{s}$ is given by

$$\mathbf{f} = \mathbf{F}(\mathbf{s}) \tag{1.5}$$

The objective of data-driven feature extraction is to estimate $\mathbf{F}(\cdot)$ so that the extracted feature $\mathbf{f}$ contains as much of the desired information as possible while suppressing the unwanted information. Note that the function $\mathbf{F}(.)$ defines all the steps involved in the feature extraction method. The following section discusses the issues involved in the proposed data-driven feature extraction paradigm.

1.5.1 Issues in Data-driven Feature Extraction

The following important issues need to be addressed in the implementation of data-driven feature extraction methods:

- **Complexity of feature extractor:** An important issue is the allowable complexity of the function $\mathbf{F}(\cdot)$ that extracts features. If the function $\mathbf{F}(\cdot)$ is unconstrained, then unreasonably large amounts of data and computation may be necessary to get a good estimate of the function. Hence in this work we use linear functions to extract features.

- **Measure of optimality:** The effectiveness of the approach depends on the optimality measure used to select the feature extraction function $\mathbf{F}(\cdot)$. If the desired source of variability is d then the ideal measure that needs to be maximized is the mutual information between the factor d and the extracted feature $\mathbf{f}$, $I(\mathbf{f}, d)$. This involves the estimation of joint probability density function of $\mathbf{f}$ and d [24, 25]. Estimation of the joint density and the following search for the optimal $\mathbf{F}(\cdot)$ is computationally cumbersome. To avoid this, we define an optimality measure so that the extracted feature $\mathbf{f}$ captures the maximum amount of variance due to the desired source while minimizing the variance captured from the undesired source. If the covariance due to the desired source of variability is Σ_d and that of the undesired source is Σ_u the linear transformation, $\mathbf{F}(\cdot)$, is estimated to maximize $\frac{\mathbf{F}^T\Sigma_d\mathbf{F}}{\mathbf{F}^T\Sigma_u\mathbf{F}}$. As seen in Section 1.2, the above measure will maximize the mutual information only if the distributions are normal with equal class-conditional covariances.

- **Generalization capability:** The ability of the estimated feature extraction method to generalize to new databases will determine the re-usability of the data-derived feature extraction module. For good generalization, the database used for estimating the feature extraction method should contain a sufficient sampling of all the possible variabilities that we expect to encounter in the application.

- **Initial representation:** The choice of the initial representation of speech from which we attempt to extract features is an important issue. To make a minimum number of *a priori* assumptions, it would be necessary to choose the speech signal itself as the initial representation. In this thesis, we use the spectrogram resulting from the short-time Fourier analysis as the initial representation $\mathbf{s}$. Hence we are not attempting to question the optimality of short-time Fourier analysis. But the general paradigm is applicable to any arbitrary initial representation including the speech waveform.

1.5.2 Advantages of the Proposed Data-driven Approach

The proposed data-driven method for feature extraction has the following advantages.

1. Feature extraction involves reduction in dimensionality. By using data-driven methods for feature extraction we are making sure that the dimensions that are preserved carry the maximum amount of useful information (or carry the least amount of harmful information) and thus improving the accuracy of the recognizer.

2. A well trained data-driven feature extraction module will suppress the unwanted variability present in the speech signal. This could potentially reduce the amount of data necessary for training the statistical classifier. This is due to the fact that the knowledge derived by the feature extraction module about the various sources of information is reused in every new task and hence need not be re-learned by the classifier. For example, if the extracted features are invariant to handset characteristics, then it is not necessary to train a speaker verification system with data from multiple handsets.

3. The data-driven feature extractor can provide knowledge about the characteristics of various sources of information carried by the speech signal. In this dissertation we will show that linear methods for feature extraction can be interpreted as spectral basis functions and temporal filters. The characteristics of these basis functions and filters can give insight into the structure of the various sources of information present in the signal.

1.6 Contributions of the Work

The dissertation is based on the hypothesis that data-driven feature extraction can provide more robust features than the conventional methods that are based on human knowledge about speech production and auditory perception. Two distinct data-driven methods are presented to support the hypothesis. In the first method, discriminant analysis is used to derive spectral features that provide better phonetic separability. The effectiveness of the discriminant spectral features is tested using ASR experiments. In the second method, data-driven techniques are used to design temporal filters to alleviate the variability introduced by telephone handsets. These filters are shown to improve the robustness of speaker verification systems. The main contributions of this work are summarized below.

- **Data-driven feature extraction:** We propose a feature extraction technique that is based on retaining the wanted source of variability while suppressing the unwanted variability. Such an approach is motivated in Section 1.5. The possible advantages of data-driven feature extraction are outlined in Section 1.5.2.

- **Discriminant spectral basis functions:** Discriminant methods are used for extracting spectral features for ASR. Spectral basis functions that preserve phonetic class separability are derived using linear discriminant analysis (LDA). For this analysis we use a large hand-labeled speech corpus that contains about 3 hours of data. The proposed discriminant features are shown to outperform the conventional features in a continuous digit recognition task.

- **Analysis of spectral basis functions:** It is observed that the discriminant basis functions analyze the low frequency part of the spectrum with higher resolution than the high frequency part. Further analysis is presented which relates non-uniform frequency resolution, phonetic discrimination, and spectral properties of speech sounds.

- **Estimation of handset variability:** A method to estimate the second order statistics of the variability introduced by telephone handsets is proposed. The proposed method uses time aligned utterances recorded using different types of telephone handsets.

- **Temporal filter design.** We introduce a data-driven method for designing filters capable of normalizing the variability introduced by telephone handsets. The design of the filter is based on the estimated second order statistics of handset variability. The filter is applied on the log energy outputs of Mel spaced filter bank. The effectiveness of the proposed handset normalizing filter in improving speaker verification performance in mismatched conditions is also demonstrated.

1.7 Outline of the Dissertation

Chapter 2 provides a brief introduction to linear transformation techniques used for feature extraction. The methods include principal component analysis (PCA) and linear

discriminant analysis (LDA). Further the chapter describes the application of PCA and LDA to derive spectral basis functions from the critical band spectrum. The spectral basis functions derived using PCA and LDA are compared to the basis functions used in the discrete cosine transform (DCT). The effect of phonetic classes on the shape of the discriminant functions are discussed. The chapter also demonstrates the capability of discriminant basis functions to reduce the error in comparison with the DCT and PCA in a continuous digit recognition task.

Chapter 3 discusses our attempt to directly derive spectral basis functions by applying discriminant analysis on DFT spectrum. This avoids the need for critical band filters. It is observed that the discriminant basis functions analyze the low frequency part of the spectrum with higher resolution than the high frequency part. The cause and effect of such basis functions are investigated. Performance of these basis functions are also evaluated by ASR experiments.

In Chapter 4, we propose a data-driven method for designing temporal filters for normalizing the variability introduced by different telephone handsets. First a brief introduction to temporal processing is provided. This is followed by an analysis of the characteristics of handset variability. This leads to a discussion of the data-driven method used for designing a handset normalizing RASTA (RelAtive SpeacTrAl) filter [23, 26]. The design of the filter is based on the estimated second order statistics of phonetic and handset variability. These estimates are used by oriented principal component analysis (OPCA) in order to derive an optimal filter. The characteristics of this filter is compared with the conventional RASTA filter [23].

The application of these filters in speaker verification is presented in Chapter 5. The speaker verification task of the 1998 and the 1999 speaker recognition evaluation conducted by the National Institute of Standards and Technology (NIST), is used for comparing the data-driven filters with conventional channel normalizing methods like RASTA and mean subtraction (MS). The effect of designing the filters using mean removed temporal trajectories are also discussed.

Chapter 6 summarizes the work and suggests future directions.

Chapter 2

Spectral Basis Function for Automatic Speech Recognition

2.1 Introduction

In the past, data-driven techniques have been used for designing temporal RASTA filters for enhancement of noisy speech [27] and for robust ASR [28, 29]. In this chapter we attempt to derive optimal spectral basis functions for ASR. Principle component analysis (PCA) and linear discriminant analysis (LDA) [30, 31] are used for designing optimized spectral basis functions for the projection of the critical-band spectrum. This is an attempt to question the optimality of the cosine basis in the context of discriminating phonemes. A brief discussion of this work can be found in [32].

The chapter is organized as follows. The next section briefly discusses the impact of using data-driven approaches on language and acoustic modeling in ASR. The section highlights the advantages of data-driven methods in comparison with knowledge based systems. This serves as a motivation for adopting data-driven methods for feature extraction. Section 2.3 reviews one of the widely used feature extraction methods, which uses an auditory-like filter-bank analysis followed by DCT. This representation serves as the baseline for all the speech recognition experiments described in this chapter. The following section discusses principle component analysis of the critical-band spectrum as a method to design a spectral basis. The characteristics of the basis function derived using PCA are compared with the conventional cosine basis. Section 2.6.3 introduces the use of LDA for deriving basis functions. The influence of the constituent phonetic classes on the shape of

the discriminant basis functions is also discussed. Results of frame based phoneme recognition experiments are reported in Section 2.7. Performance of the discriminant basis functions is further analyzed by connected digit recognition in Section 2.8.

2.2 Data-Driven Approaches in ASR - A Brief History

Typical large vocabulary automatic recognition of speech (ASR) consists of three main components: feature extraction, pattern classification, and language modeling. Feature extraction attempts to reduce the data rate of raw speech data by alleviating irrelevant variability such as speaker characteristics or environmental noise. Pattern classification further reduces the data rate by classifying each time instant into one of the subword-unit classes, and language modeling compensates for possible errors of classification by emphasizing more likely word combinations.

Early ASR systems were inspired by the advances in artificial intelligence (AI) [33]. These systems relied on sets of rules for pattern classification and language modeling. These rules were often prescribed by *experts* in reading spectrograms and hand crafted for specific recognition problems at hand [34, 35, 36]. They worked reasonably well for small tasks under controlled environments. The performance of such systems were found to be fragile [37].

The past two decades have witnessed the introduction of stochastic approaches in both the pattern classification and the language modeling modules. These stochastic approaches brought the rich mathematical basis that was available in the classical pattern recognition literature to ASR. In the current ASR systems, the pattern classification module uses Hidden Markov Models (HMM) [19, 2] and Artificial Neural Networks [38] while the language models are in the form of N-grams which are trained from a large text corpus [39]. Stochastic techniques typically use only minimal *a priori* assumptions about the nature of the problem. Such techniques estimate the parameters of models directly from the data. Replacing the hardwired prior knowledge by the knowledge derived from the data turned out to be one of most significant advances in ASR research.

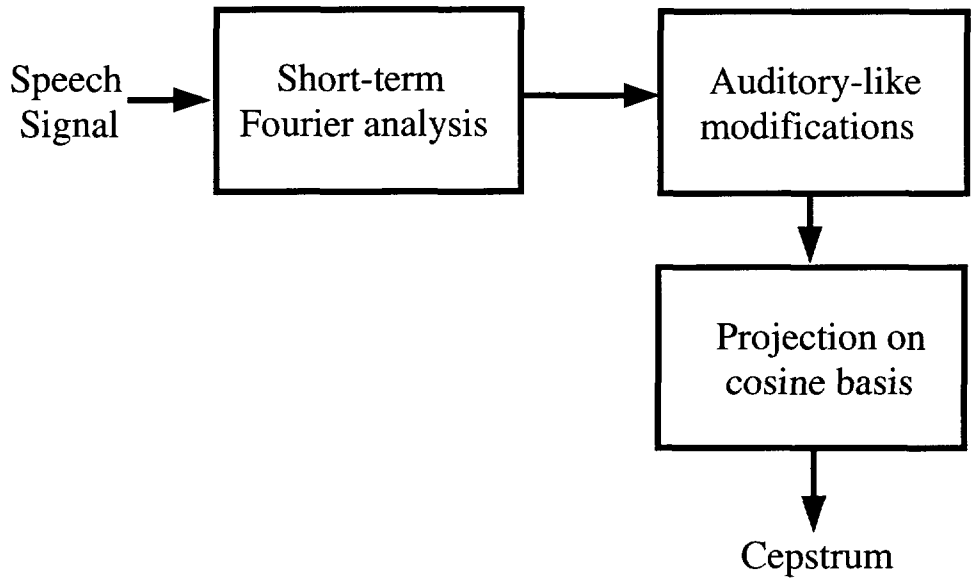


Figure 2.1: Generic form of dominant speech representations used in ASR. The short-time speech spectrum is modified by auditory-motivated processing. This may include warping of the frequency axis and smoothing of the spectrum by simulated auditory-like filters. Modifications along the amplitude axis typically involves a logarithmic nonlinearity. The modified spectrum is then projected on cosine basis.

2.3 Current Feature Extraction Techniques for ASR

Data-driven approaches are still largely absent in the analysis module. Only recently, some emerging efforts in deriving temporal RASTA processing from the data have started to appear[28, 29, 40]. The current work attempts to extend such data-driven techniques into optimization of spectral bases in speech analysis.

The analysis module in ASR typically consists of a series of processing steps as shown in Figure 2.1. Some of these steps are inherited from speech coding, and some justified by perceptual or pattern matching arguments. A widely used speech representation is the auditory-like cepstrum [12, 14]. This cepstrum represents an appropriately modified (through auditory-like frequency and amplitude warping and critical-band smoothing) short-term spectrum of speech, projected onto a cosine basis. The short-term spectrum is derived from about 25 ms long consecutive segments of the speech signal. The spectral

modifications are justified by properties of human hearing [14], and the cosine projection by the need for partial decorrelation of features [41] used in the subsequent pattern classification.

2.3.1 Short-time Fourier Analysis

Short-time analysis is one of the most widely used and accepted processing techniques for speech. The fundamental assumption in short-time analysis is that even though speech is non-stationary, over sufficiently short-time intervals it can be considered stationary. Hence the Fourier transform of a short segment of speech should give a good spectral representation of speech. The discrete short-time Fourier transform of a signal $s(n)$ is given by

$$S(\omega, t) = \sum_{n=t-k}^{t+k} s(n)h(t-n)e^{-j\omega n},$$

where t is the discrete time index and $h(t)$ is a symmetric window applied to the short-time signal and is non-zero only in the interval $-k$ to k. Typically a Hamming window of 25 ms duration is used. Such a window when applied on a signal which is sampled at 8 kHz will band-limit $S(\omega, t)$ to 80 Hz in t dimension. Hence $S(\omega, t)$ needs to be sampled at 160 Hz [42, 43]. However, in practice, we have observed that sampling at 100 Hz (every 10 ms) makes no difference to the processing.

2.3.2 Non-Uniform Frequency Resolution

The squared magnitude of the short-time Fourier spectrum, $S_p(\omega, t) = |S(\omega, t)|$, is further processed to simulate the non-uniform frequency resolution observed in human auditory perception. The notion of critical bands suggests that the resolution of the human ear is non-uniform across the frequency band [44]. This is due to integration of energy of signals that fall within a critical band. Psychoacoustic experiments using simultaneous frequency masking have revealed that the bandwidth of critical bands increases with the center frequency. This property of hearing is simulated in the current feature extraction modules by projecting the short-time power spectrum onto the frequency responses of the simulated critical band filters, represented by the following equation

$$s_{l,k} = \ln[\mathbf{f}_k^T \mathbf{s}_p] \qquad k = 1, 2 \ldots 15, \tag{2.1}$$

where $\mathbf{f}_k$ and $\mathbf{s}_p$ are column vectors representing the frequency response of the kth critical bands and the short-time power spectrum respectively. This operation yields $s_{l,k}$ which is the logarithmic energy output from the kth critical band.

2.3.3 Projection onto Cosine Basis

The logarithmic critical band spectrum is then projected onto a cosine basis by the matrix multiplication,

$$\mathbf{c}_c = \mathbf{C}\mathbf{s}_l \tag{2.2}$$

where $\mathbf{s}_l$ is the vector of 15 critical band energies and $\mathbf{C}$ the matrix whose rows are the cosine basis function. The feature vector $\mathbf{c}_c$ resulting from the DCT is also known as the cepstrum[1]. It has been shown that for vectors which are generated by a first order Markov processes, projection onto a cosine basis approximately decorrelates the distribution [41, 47, 48]. Decorrelated feature vectors can be effectively used in training HMMs that use diagonal covariance matrices for modeling class conditional distributions. The capability of the DCT to decorrelate the features is illustrated Figure 2.2. The figure shows the covariance of the critical band spectral vectors computed over the English portion of the OGI STORIES corpus. This portion of the corpus consists of fluent telephone-quality speech from 210 adults of both genders. It is clear from the figure that the cosine basis functions make the total covariance close to diagonal. But still the covariance matrix computed in the cepstral space does have significant off diagonal elements. It is also interesting to note that the variance captured by the first cosine basis vector amounts to

[1]Typically only the first few components of this cepstral vector are used for ASR. This truncation leads to a smoothing of the auditory spectrum. One of the first uses of cepstrum was in the form of homomorphic deconvolution. Truncation of a cepstral vector leads to what is known as cepstral smoothing of the log-magnitude spectrum, thus deconvolving the source and the system characteristics present in the log-magnitude spectrum. This technique, to our knowledge, was first introduced in speech processing by A. M. Noll [45] and later extended for spectral smoothing by A. Oppenheim [46]. In the case of auditory power spectrum this smoothing has already been achieved up to some extent by the critical band filters.

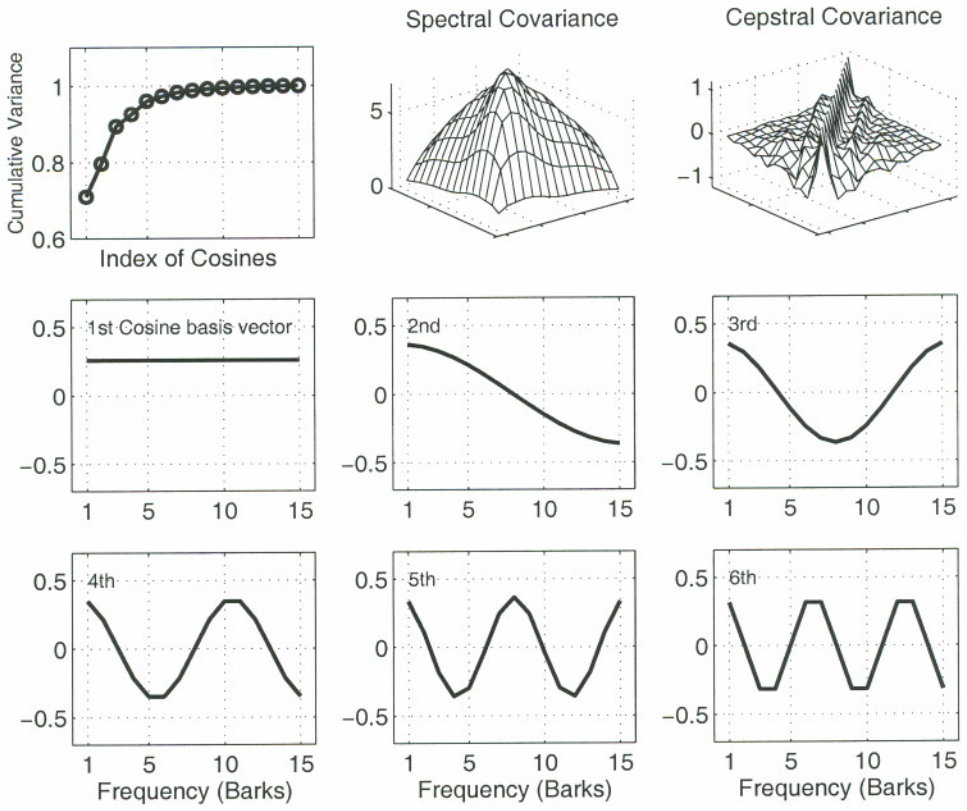


Figure 2.2: Upper left: The cumulative variance captured by using progressively higher frequency cosine basis vectors. Upper center: Covariance matrix of the original critical-band auditory spectral space derived from about 3 hours of OGI STORIES speech corpus. As seen, the spectral covariance matrix is far from diagonal. Upper right: The total spectral covariance matrix, projected on the first 8 vectors of the cosine basis, is partially diagonalized. The figure also shows the first 6 cosine basis functions.

about 70% of the total variability. Since the first cosine basis vector is flat across the critical bands, it means that the majority of variance in the speech spectrum is caused by variation in average energy. The first 10 cosine basis functions capture almost the entire variance present in the data. This indicates that the 15 dimensional critical band spectrum can be represented by a significantly lower dimensional vector which supports the cepstral truncation often performed in ASR. This also indicates the amount the redundancy that is present in spectral vectors.

The rest of the chapter explores alternative linear mapping techniques to the DCT for

improving the performance of ASR. The technique used is based on analyzing spectral features extracted from about 3 hours of speech data from the OGI STORIES corpus. This corpus thus contains all major sources of variability encountered in the speech signal, including channel/handset variability, context/coarticulation variability, speaker variability, and phonetic variability.

2.4 Linear Mapping for Feature Extraction

This section briefly reviews linear mapping techniques for extracting features used in this as well as the following chapter for extracting features. Principal component analysis (PCA) and linear discriminant analysis (LDA) are two linear mapping techniques which are extensively used in pattern recognition. While principal component analysis focuses on preserving as much variance as possible, LDA tries to preserve the variability that maximizes linear separability of 2 or more classes. Both PCA and LDA are linear techniques that try to extract features from a random vector by projecting the vector onto a set of *basis vectors*. Consider an N dimensional random vector $\mathbf{x}$ having zero mean. Then a set of features y_i can be extracted from $\mathbf{x}$ by projecting $\mathbf{x}$ onto feature-extracting basis functions given by

$$y_i = \mathbf{x}^T \phi_i \qquad i = 1, 2 \ldots N. \tag{2.3}$$

In the above equation, ϕ_i is the basis which extracts features by linearly combining (weighted average) the components of the random vector $\mathbf{x}$. Note that, for each of the features y_i $i = 1, 2 \ldots N$ to carry unique information about $\mathbf{x}$, the basis vectors, $\phi_i, i = 1, 2 \ldots N$ have to be linearly independent. It is the properties of ϕ_i that differentiate PCA from LDA, and the next two subsections elaborate on these properties. Also note that the DCT is a special case of a linear mapping where the basis functions, ϕ_i, are cosines.

2.5 Spectral Basis from Principal Component Analysis

2.5.1 Principal Component Analysis

If the basis functions ϕ_i in Equation 2.3 are the eigenvectors of the covariance matrix of $\mathbf{x}$, then the resultant feature extraction is termed as principal component analysis [49]. This is also the discrete version of Karhunen-Loéve transform (KLT). The basis functions ϕ_i are used to represent $\mathbf{x}$ with an M dimensional vector where $M < N$. The PCA-derived basis guarantees that the lower-dimensional representation gives minimum reconstruction error. Since ϕ_i are orthogonal to each other[2], $\mathbf{x}$ can be reconstructed from y_i using the following equation.

$$\hat{\mathbf{x}} = \sum_{i=1}^{M} y_i \phi_i. \tag{2.4}$$

Note that ϕ_i are sorted according to the corresponding eigen-values. Hence, if $\lambda_i, i = 1, 2 \ldots N$ are the corresponding eigen-values, then $\lambda_i \geq \lambda_j$ if $i < j$. If the eigenvectors of the covariance matrix of $\mathbf{x}$ are used as ϕ_i then the error in reconstruction, ε, given by

$$\varepsilon = E\left[(\mathbf{x} - \hat{\mathbf{x}})^2\right],$$

is minimum for any M. It can also be shown that the covariance matrix of the vector $\mathbf{y} = [y_1 y_2 \ldots y_N]$ is diagonal. Hence the two main advantages of using the PCA basis to extract features are that (a) Dimensionality reduction guarantees minimum reconstruction error, and (b) the covariance matrix of the extracted feature is diagonal, aiding the use of simpler models. Note that PCA is designed to minimize the reconstruction error and not the classification error.

2.5.2 Principal Component Analysis of Critical Band Spectrum

In this section, the effectiveness of the cosine transform in decorrelating the speech representations is studied. This is done by comparing the cosine basis to the basis function

[2]Since covariance matrices are symmetric, eigenvectors of a covariance matrix are always orthogonal to each other.

obtained from principal component analysis which involves solving the following eigenvalue problem,

$$\Sigma \mathbf{e}_i = \lambda \mathbf{e}_i. \tag{2.5}$$

where Σ is the covariance matrix of the feature vectors extracted from the STORIES database. The principal component analysis was performed on 15-dimensional critical band spectrum obtained by using Bark spaced trapezoid-shaped filters [14]. Figure 2.3 shows the basis functions derived through PCA of critical band spectrum. The basis vectors are reminiscent of cosine functions, with the first vector measuring the spectral energy, and the consecutively higher ones similar to cosine-like functions with decreasing period. The transformed covariance matrix is of course diagonal.

The following two conclusions are drawn from this analysis:

- The DCT makes the total covariance of the critical band spectrum approximately diagonal.
- It is clear by comparing Figure 2.3 and Figure 2.2 that the PCA basis is very similar to the cosine basis. Thus, DCT is a good choice for decorrelating the critical band spectrum. This also indicates that cepstral truncation can be used as a method to reduce dimensionality, which results in minimum reconstruction error.

2.6 Discriminant Analysis of Critical Band Spectrum

2.6.1 Linear Discriminant Analysis

In speech recognition, the features extracted from the signal are used to classify the sounds into phonetic categories. Hence a feature extraction technique should be designed to preserve as much class separability as possible. An ideal feature extractor should be able to reduce the error to its theoretical limit, which is given by Bayes' error [49]. For an L class problem, the Bayes' classifier that yields minimum error compares L a posteriori probabilities, $p_{1|\mathbf{x}}(\mathbf{x})$, $p_{2|\mathbf{x}}(\mathbf{x})$ $\ldots p_{L|\mathbf{x}}(\mathbf{x})$, and classifies $\mathbf{x}$ to the class that gives maximum *a posteriori* probability. Another interesting point is that the a posteriori probabilities are

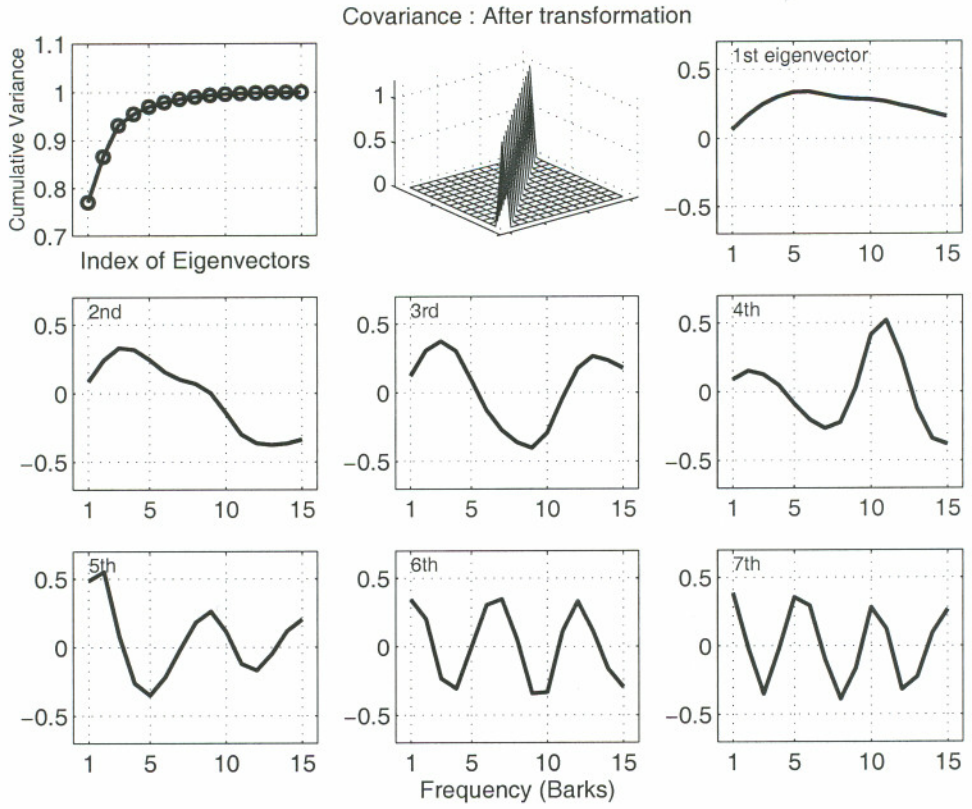


Figure 2.3: Upper left: The cumulative variance captured by the first 8 principal components. Upper center:The total spectral covariance matrix projected on the basis vectors given by principal component analysis. The first six basis functions derived by PCA of the critical-band spectral space are also shown.

not linearly independent, since

$$\sum_{i=1}^{L} p_{i|\mathbf{x}}(\mathbf{x}) = 1. \tag{2.6}$$

Only $L-1$ of the L posterior probabilities are linearly independent. Hence these $L-1$ linearly independent features are the ideal set of features that would give an error which is equal to the error given by a Bayes' classifier. That is, by transforming the original random vector $\mathbf{x}$ into an $L-1$ dimensional vector by using the transformation, $y_i = p_{i|\mathbf{x}}(\mathbf{x})$, $i = 1, 2, \ldots L-1$, no classification information is lost.

Even though Bayes' error is the best criterion and the *a posterior* probabilities the best features, *a posteriori* probability functions are hard to estimate without severe biases

and variances. Hence, in discriminant analysis simpler measures are used to optimize the feature extraction process.

The basis functions used for LDA are designed to maximize linear separability between classes. Linear separability is defined with the help of within-class and between-class scatter matrices (covariance matrices). The within-class covariance represents how much the samples within a class vary and is given by

$$\begin{aligned} S_w &= \sum_{i=1}^{L} P_i E\left[(\mathbf{x}-\mu_i)(\mathbf{x}-\mu_i)^T|\omega_i\right], \\ &= \sum_{i=1}^{L} P_i \Sigma_i, \end{aligned} \tag{2.7}$$

where μ_i and P_i are the class conditional mean and the *a priori* probability of the i^{th} class respectively. The between-class covariance matrix is the covariance of the class conditional means themselves given by

$$\begin{aligned} S_b &= \sum_{i=1}^{L} P_i \left[(\mu-\mu_i)(\mu-\mu_i)^T\right], \\ \mu &= \sum_{i=1}^{L} P_i \mu_i, \end{aligned} \tag{2.8}$$

where P_i is the *a priori* probability of class i.

A property of S_w and S_b is that they sum up to the total covariance and hence LDA can be viewed as decomposing the total variability into within-class and between-class variability. A widely used criterion for class separability is defined by

$$F = trace(S_w^{-1} S_b). \tag{2.9}$$

It is straight forward to show that the set of basis vectors ϕ_i, that satisfies the following generalized eigen-value problem

$$S_b \phi_i = \lambda_i S_w \phi_i, \tag{2.10}$$

maximizes F. It must be noted that the matrix S_b can have a maximum rank of $L-1$. Which in turn makes the rank of $[S_w^{-1} S_b]$ also $L-1$. Hence the maximum number of linearly independent basis functions that can be computed using the above equation is

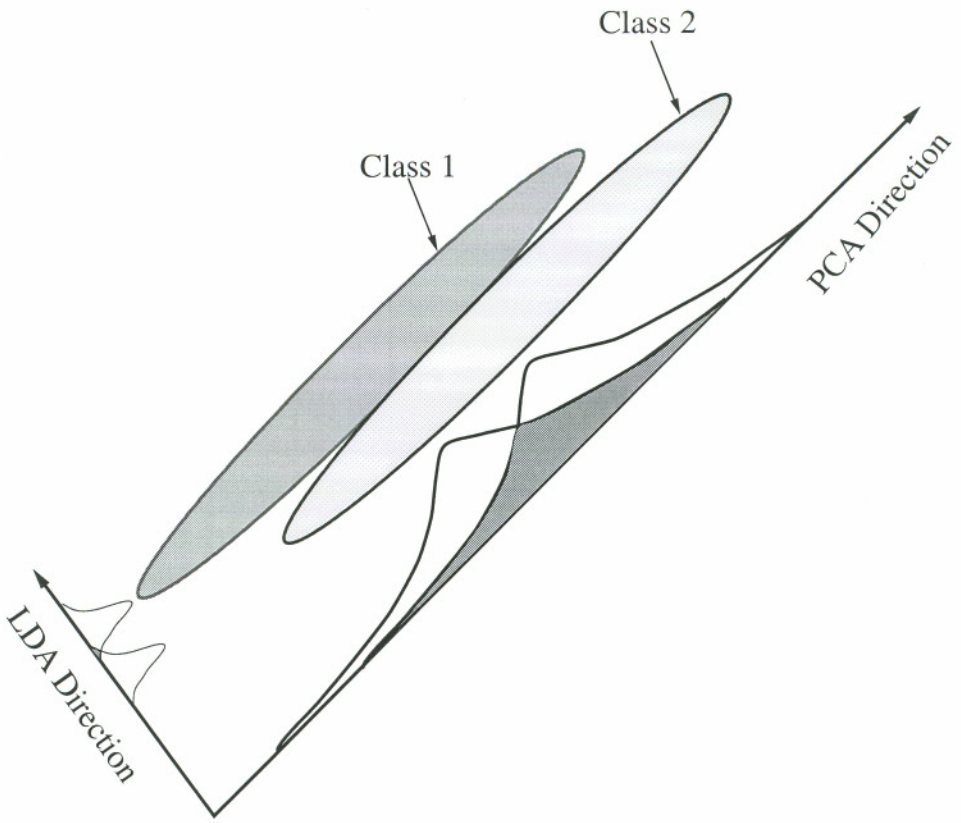


Figure 2.4: Illustration of the difference between PCA direction and LDA direction. Projecting the original two-dimensional vector onto the PCA direction preserves maximum variability. But the resultant class conditional distributions overlap significantly. On the other hand, projection onto the LDA direction yields a one-dimensional feature with minimal overlap of class conditional distributions.

$L-1$. Figure 2.4 highlights the difference between LDA and PCA. From the figure it is clear that PCA may provide features which result in significant overlap of class conditional distributions, which in turn results in higher error. On the other hand LDA yields features with minimum overlap of class conditional distributions.

It should be noted that the basis functions computed using LDA are optimal only under the following assumptions:

- The class conditional distributions are normal.
- All the class-conditional covariances are equal.

- Only two classes are involved in the classification.

If the above conditions are met, then the basis functions given by LDA are guaranteed to be optimal and the classification error will be equal to the Bayes' error bound. When these conditions are not met, dimensionality reduction using discriminant basis vectors can increase the overlap of class conditional distributions. This in turn causes an increase in the lowest theoretical achievable error. But practically, even for multi-modal distributions, $trace(S_w^{-1}S_b)$ has been found to be a good measure for class separability unless the class conditional means are all the same [49]. Moreover, in this thesis, LDA is used not only as a feature extraction method to reduce classification errors, but as a analysis tool to understand the nature of variability that is important for phonetic discrimination.

2.6.2 Review of Discriminant Methods for Feature Extraction in Speech

LDA is not new to speech processing. To our knowledge, its use has been first studied by Hunt [41]. Later, in a series of articles, Hunt and his colleagues used LDA to combine disparate acoustic parameters including temporal derivatives to improve performance and for dimensionality reduction [50, 51, 52]. They call it IMELDA - Integrated Mel-scale representation with LDA. In all these studies, classes defined by the states of HMMs were used to compute the within- and between-class covariances. These statistics were derived from trained HMMs representing words. More recently Ayer extended the idea of IMELDA to optimize word separability using gradient decent methods, which he calls whole-word adaptive LDA (WALDA) [53]. This version of LDA was designed to focus on those word tokens that yield borderline recognition decisions. Discriminant analysis was also used by Brown to process several concatenated feature vectors, thus addressing both temporal and spectral dimensions [54].

Doddington uses a state specific discriminant transformation for improving the recognition performance. For each of the HMM states a transformation which enhances the discrimination between that state and other *confusable* states is derived [55]. A similar method for improving the recognition performance of an HMM based ASR system was reported by Woodland and Cole [56].

Linear discriminant analysis is further optimized by Schukat-Talamazzini in a maximum likelihood frame-work and is used to optimize the projection used for feature extraction and HMM parameters simultaneously [57]. However this method was found to be impractical due to its computational needs. Finally a comparison of various linear transformations as alternatives to the DCT can be found in [58, 59].

All the work related to the application of discriminant analysis found in the literature focuses on improving the recognition performance of a specific task by using discriminant vectors instead of the DCT. In this thesis (in the current and following chapters), the focus is on deriving discriminant vectors from various initial representations. The objective is not only to improve the recognition performance but also to understand the characteristics of the short-time spectrum relevant for phonetic discrimination. The work reported in this dissertation regarding the use of discriminant linear transformation is unique in the following aspects:

1. We use phonemes as classes and relatively large labeled databases to estimate S_w and S_b and not the classes defined by states in HMM . This makes the results less dependent on the particular HMM structure used for recognition.

2. For estimating the LDA basis functions, we use a sufficiently large database (approximately 3 hours of speech) which contain all the major sources of information. The effectiveness of the discriminant basis functions are then evaluated by ASR experiments. We consistently use an independent corpus for this evaluation. This would help in evaluating the generalization capability of the proposed discriminant analysis.

3. We analyze and interpret the shapes of the discriminant basis functions. Such an analysis, for example, could lead to the understanding of the type of spectral changes which carry significant phonetic information.

4. We apply discriminant analysis directly to the DFT-spectrum, bypassing the filter-bank analysis. This leads to establishing a link between nonuniform frequency resolution (implied by Mel and critical-bank integration) and phonetic discrimination.

In the next section, we investigate the use of LDA to derive spectral weighting functions (spectral basis) as an alternative to the cosine basis used in the conventional Mel/critical-band cepstral analysis.

2.6.3 Discriminant Basis Functions

From the previous section it is clear that the DCT approximately decorrelates the features. In this section, the optimality of such a rotation in preserving phonetic discriminability is questioned. In this context it is assumed that phonemes are the basic units for speech recognition. Hence the rotation and dimensionality reduction should be able to preserve the variance introduced by phonemes while suppressing the variance introduced by sources like coarticulation, channel and speaker. The PCA/DCT basis projects the features onto the directions of maximum variability. As known, there are many sources of variability in speech, many of them harmful for phonetic classification [60, 61, 62]. Hence it is desirable to project the space on the direction of maximum separability rather than on the direction of maximum variance. As described in the previous section linear discriminant analysis estimates the direction of maximum linear separability.

As seen in the previous subsection, computing the discriminant vectors involves generalized eigendecomposition of (S_b, S_w), where S_w refers to the matrix of the mean of the covariances of classes and S_b the matrix of the covariance of the means of the classes. Figure 2.5 shows the diagonal elements of the within- and the between-class covariance matrix, computed on critical band spectral features from the STORIES database. This computation involved approximately 650,000 feature vectors. It is interesting that the between-class variance is significantly higher than the within-class variance at around 5 barks. This suggests that the maximum amount of discriminable information is present at around 500 Hz. This observation is consistent with the work reported by Yang et. al., [24] where it was shown that the mutual information between phonetic labels and spectral measurements is the highest for the spectral band centered at around 500 Hz. Figure 2.5 also illustrates the effect of removing the mean of the spectral vectors from each utterance on within- and between-class variance. It is clear that mean normalization significantly reduces the within class variance without affecting between-class variance. This can be

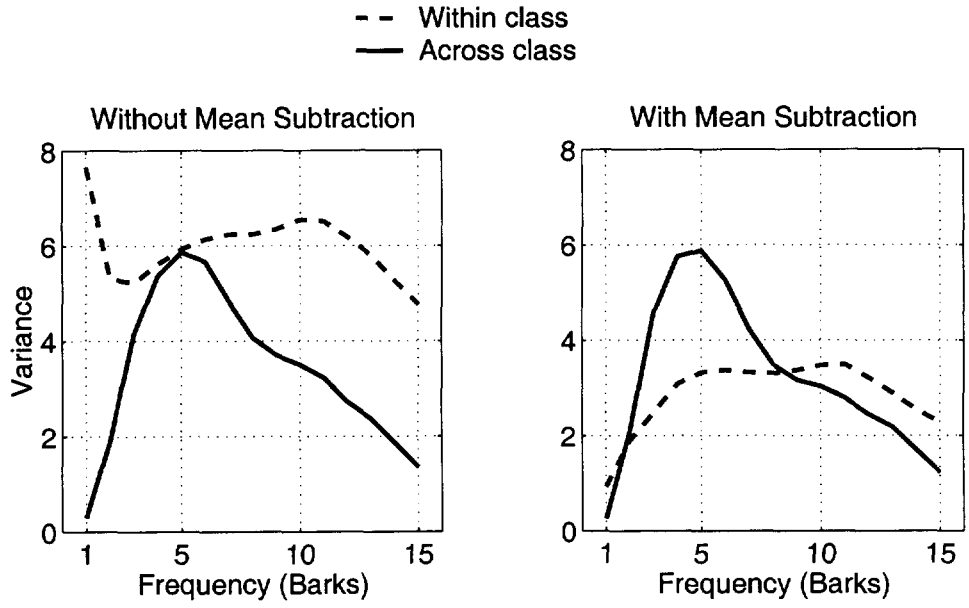


Figure 2.5: The within- and the between-class variances computed on critical band spectral feature with and without utterance based mean subtraction.

explained as follows: Long-term average of logarithmic spectra primarily contains information about the speaker and channel [62], both of which contribute to within-class variance. Hence subtracting the average spectral vector from each of the utterances reduces the within-class variance without affecting the between-class variance.

We have often observed that a direct computation of the discriminant eigenvectors leads to noisy solution. Such a solution is characterized by discriminant vectors that are not smooth. This is because of directions in the feature space that yield a high F-ratio but contain very little variance[3]. In this work we alleviated this problem by first smoothing the critical-band energy vectors using truncated PCA, which preserved 99% of the original variance in the data. This eliminates directions in the feature space containing less than a 1% of total variance and makes the solution relatively noise free. The result is illustrated in Figure 2.6 which shows the first six eigenvectors (linear discriminants) of the PCA-smoothed (99% of variance) Fisher discriminant matrix [63].

[3]Since F-ratio is the ratio of between- and within-class variance, it can assume a large value even if both the between- and within-class variances are small.

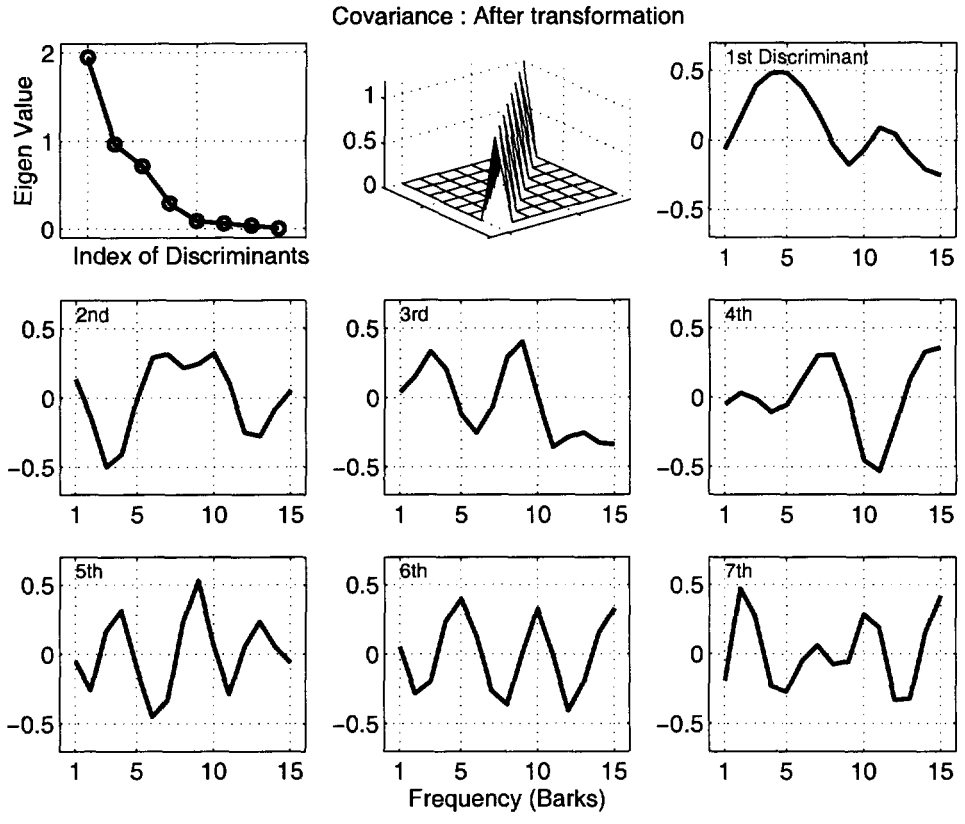


Figure 2.6: Upper left: Eigenvalues of the LDA-derived basis vectors. Upper center: The total critical-band spectral correlation matrix, projected on the first 8 basis vectors of the LDA-derived basis. The first 7 LDA-derived spectral basis functions of the critical-band spectral space are also shown.

The first four discriminant basis functions seems to be significantly different from cosine functions. Unlike the first cosine function, the total energy of the spectrum is no longer measured. The first discriminant appears to evaluate spectral energy in the first formant region and could be primarily discriminating between sonorant and non-sonorant sounds. The second and third discriminants seem to be focusing on spectral ripples [64] in the central part of the critical-band spectrum, the second one being more sensitive to larger ripples than the third one. The 4th one analyses the portion of the spectrum that lies above 5 Barks. The 5th discriminant vector is sensitive to spectral ripples with a 5 Bark period. The 5th and 6th discriminants are very similar to sinusoidal functions.

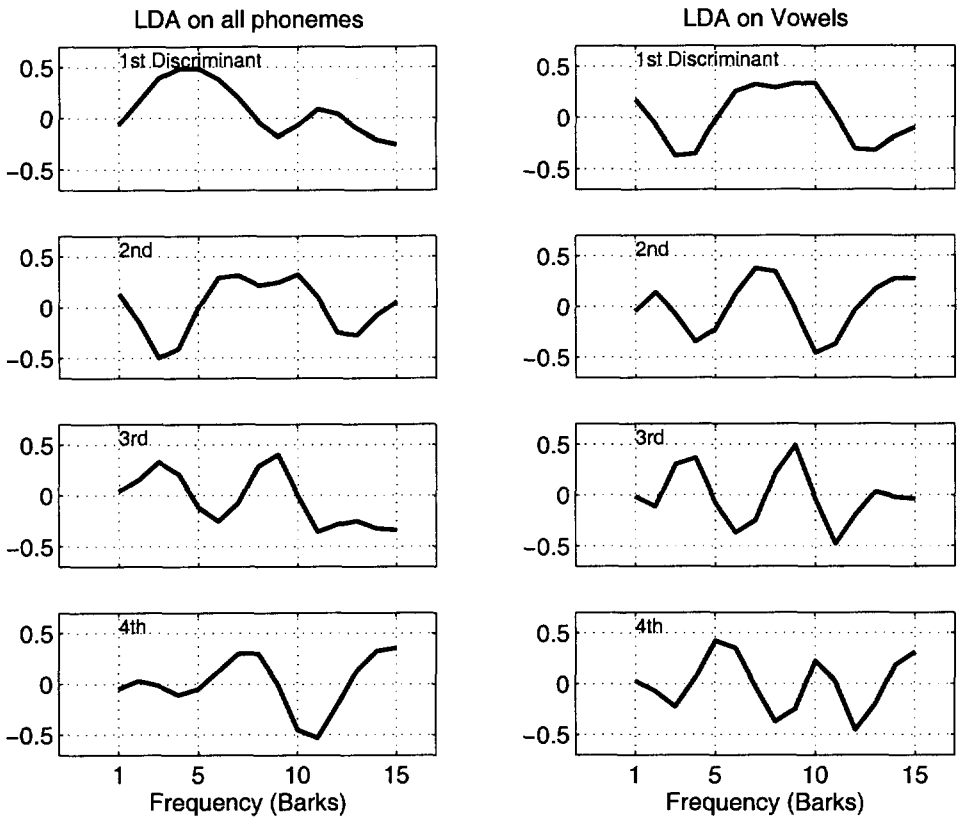


Figure 2.7: Effect of constituent phonetic classes on the shape of discriminant basis vectors derived using LDA.

The higher discriminants are less important as the eigenvalues associated with them are relatively small. Just as in the PCA, the LDA-transformed covariance matrix is also diagonal.

2.6.4 Effect of Constituent Classes on the Discriminant Vectors

It is important to note that the shape of the discriminant spectral basis vectors is determined by the phonetic classes under consideration. In order to illustrate this point the design of the discriminant spectral basis functions was carried out by considering only the vowels. In Figure 2.7, the resultant discriminant basis functions are compared with those computed by LDA of all phonetic classes. The major difference between the two sets of basis is in the first discriminant vector. When all the phonemes are considered

the first discriminant vector seems to integrate the energy in the first formant region as well as analyze the spectral slope of the spectrum. Such an analysis could only be helpful in distinguishing sonorant sounds from non-sonorants. Hence the vectors designed to discriminate various vowels does not have such a function. Another difference is in the absence of frequency selective basis function (4th).

The dependency of the discriminant functions on the set of phonemes involved in the design suggests the use of a different spectral basis depending upon the vocabulary used in the ASR task.

2.7 Phoneme Classification Experiments

To asses the effectiveness of the data-derived spectral basis we ran a phoneme-classification experiments. T he task was to classify individual frames of the test set into one of the 29 phonemes. The classification was based on a single spectral frame. Logarithmic spectral mean was subtracted from each file to partially compensate for communication channel differences. Speech from about 1400 files from the OGI NUMBERS corpus were used in the training of a simple single-density, diagonal-covariance Gaussian classifier. Each file contains an utterance of a sequence of digits (zip codes, telephone numbers etc.) spoken by a single speaker. Approximately 120, 000 spectral frames from around 800 utterances were used for testing. This experimental setup was used for evaluating the following three different spectral bases vectors.

1. Cosine spectral basis (8 DCT coefficients).

2. Spectral basis from PCA of OGI STORIES corpus (first 8 coefficients).

3. Spectral basis from LDA of OGI STORIES corpus (first 8 coefficients).

Results of these experiments are shown in Table.2.1. The full covariance classifier results are practically identical (they should be identical if there was no truncation of higher basis functions since such classifier is invariant under linear projections). For the diagonal covariance case, which is of more interest for HMM classification, the LDA-derived basis vectors perform better than both the cepstrum and the PCA basis. This difference in

Table 2.1: Phoneme classification error on the OGI Numbers corpus

Covariance	DCT	PCA	LDA
DIAG	53.10 %	54.00 %	52.10 %
FULL	51.30 %	51.32 %	51.27 %

performance is significant according to the binomial test for a chosen significance level of 0.001 (i.e at 99.9% confidence level).

2.8 Connected Digit Recognition Experiments

In order to asses the usefulness of the discriminant basis functions in a practical ASR application these basis functions were used as feature extractors in an HMM based connected digit recognizer. The task is to recognize connected digits spoken by multiple speakers over different telephone channels. The digits part of the OGI numbers database was used for this purpose. The training part consists of about 2500 training utterances and the testing part consists of around 2200 utterances. Context independent monophones were the basis acoustic models for decoding the digit strings. Multiple pronunciations for the digits were also considered.

Figure 2.8 shows the word error rate for the discriminant features as compared to the DCT and PCA features. The error given by the discriminant features is consistently lower than the PCA and DCT features. For lower dimensional features, LDA features yield significantly lower error. While using 4-dimensional features the error reduction is more than 30%. Hence it can be concluded that by using a discriminant spectral basis, we can either achieve lower error, or we can maintain the same performance with smaller number of features.

The performance of the LDA-derived feature vectors with additional dynamic features given by velocity and acceleration coefficients [65] was also analyzed. Table 2.2 summarizes the results. For all these experiments the delta and acceleration coefficients were appended to the DCT- and LDA-derived feature vectors. For these experiments the architecture of the HMMs was varied to observe the effect of complexity of modeling (number of

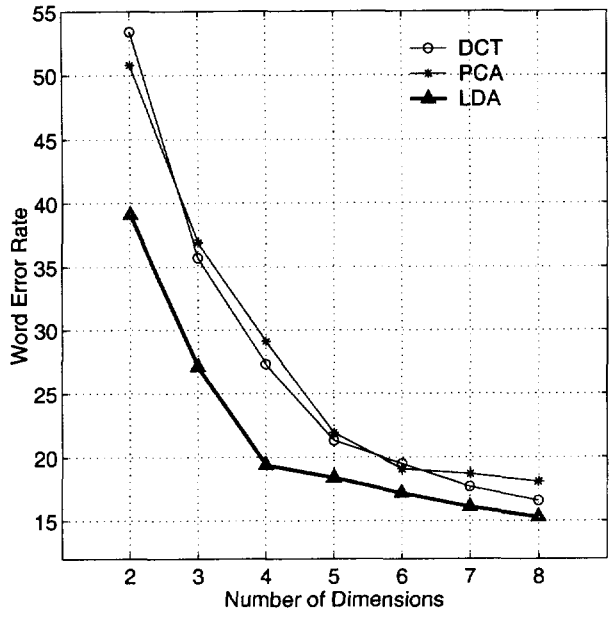


Figure 2.8: Word error rate given by PCA, DCT and LDA features as a function of the number of dimensions of the feature space.

parameters used to model the distribution of the feature vectors). From the table it is

Table 2.2: Comparison of word error rates given by cosine (DCT) basis function and discriminant functions derived using LDA. The task is to recognize continuously spoken digits from the OGI Numbers corpus. For the above experiments, 8 features appended with the velocity and acceleration coefficients were used.

HMM Structure	DCT	LDA
1-state 1-mixture	29.6 %	28.3 %
3-state 3-mixture	7.7 %	7.0 %
5-state 3-mixture	6.5 %	5.8 %
5-state 5-mixture	5.7 %	5.0 %

clear that irrespective of the complexity of modeling, the LDA basis functions outperform the DCT. It must be noted that only the 1-state 1-mixture model matches with the assumptions that went into LDA while in other cases mixture models with multiple states were used to model the class conditional distribution. Hence it is very encouraging to note that even at very low error rates (from 5-state 3-mixture case) when the complexity of

modeling is very high the LDA basis performs significantly better than the DCT.

2.9 Discussion and Conclusions

In this chapter we first examined the nature of the basis functions that correspond to the direction of maximum variability derived using principal component analysis of the critical band spectrum. The results indicate a high similarity between these basis functions and the cosine functions used by DCT. Linear discriminant analysis (LDA) was performed on the critical-band spectrum by considering context independent phonemes as classes. The discriminant basis functions were found to be significantly different from the cosine functions. The alternative spectral basis functions offer an advantage in phoneme classification of spectral vectors. This advantage seems to hold irrespective of the complexity of the task. The LDA derived basis functions outperformed the conventional DCT in connected word recognition tasks as well.

Periodicity of the optimized spectral bases in the Bark domain could point to the usefulness of Bark-like spectral warping in phoneme classification. The zero-crossings of the LDA-derived spectral basis functions are reasonably uniformly spaced on the Bark scale of the auditory-like critical-band spectrum. Thus, it appears that the Bark frequency scale allows for use of simple basis functions in phoneme classification. The following chapter further examines the relevance of Mel/Bark warping in the context of phonetic discriminability.

Chapter 3

Non-uniform Frequency Resolution from Speech Data

3.1 Introduction

In the previous chapter the optimality of the DCT applied to the critical-band spectrum was analyzed. It was shown that basis functions derived using linear discriminant analysis perform better than the DCT. In this chapter the utility of critical-band analysis itself is investigated. Use of critical-band analysis (or Mel filter-bank analysis) is motivated by properties of hearing. This analysis provides a non-uniform resolution to the short-term Fourier spectrum of speech signal. In this chapter we investigate the usefulness of such an analysis from the point of view of pattern recognition and spectral properties of speech sounds. Through such an analysis we show that the Bark/Mel like frequency resolution automatically results from discriminant analysis of short-time Fourier spectrum of speech. We also show that such a resolution can be traced to the spectral properties of vowels. These properties are further related to the physiology of the vocal tract.

This chapter is organized as follows. The next section briefly discusses the origin of the concept of critical bands as it evolved out of perceptual experiments. It also describes the signal processing methods used to emulate this property of hearing. Section 3.3 reports the results of the discriminant analysis performed on the DFT spectrum and discusses the nature of discriminant information across the spectrum. The effect of constituent phonetic classes on the nature of discriminant functions is the focus of Section 3.4. Section 3.5 explores the causes for the non-uniform resolution found in the spectral discriminant

function. The effect of such non-uniform resolution in terms of sensitivity to formant location is discussed in Section 3.6. ASR experiments with features extracted by discriminant basis functions are reported in Section 3.7.

3.2 Non-uniform Spectral Resolution - Its Origin

3.2.1 Critical Bands

The presence of one sound may obscure or even prevent the perception of another sound. The process by which the threshold of audibility of a sound is raised by the presence of another sound is termed as *masking*. Perceptual experiments involving masking can provide useful information about properties of hearing. In this section, earlier work involving such perceptual experiments is reviewed. Specifically of interest is *simultaneous masking* in which the signal is presented simultaneously with the masker. It was these experiments that led to the concept of critical bands.

The first systematic study of the masking of a pure tone by another was reported by Wegel and Lane [66]. In their work, the threshold of audibility of a tone with varying frequency was estimated in the presence of a fixed frequency sinusoidal masker. The masked audiogram[1] generated from these experiments show steep slopes on the low frequency side. The slopes on the high frequency side are less steep. The notion of critical bands was introduced by Fletcher [67] in order to explain many such phenomena related to simultaneous masking. In particular it was devised to explain the masking phenomenon of a narrow band (sinusoidal) signal by wide-band noise. He hypothesized that the early auditory system works as a bank of bandpass filters with overlapping center frequencies. He again hypothesized that only a narrow band of frequencies surrounding the tone (the critical band) contributes to the masking of a pure sinusoid. This hypothesis was later confirmed with the help of experiments in which the threshold of hearing a pure tone was measured in noise. It was observed that an increase in noise bandwidth beyond a certain limit does not effect the threshold [68, 69]. Figure 3.1 shows the critical bandwidth as a

[1]The figure showing the variation of the threshold of audibility as a function of the frequency of the masker is called the *audiogram*

function of the center frequency. The estimation of critical bandwidth is due to Zwicker *et al.*, [44]. It is clear from the figure that the critical bandwidth increases with frequency. This suggests that the ear has more resolution in the lower frequency region than in the high frequency region. A low frequency pure tone is masked only by sounds that are

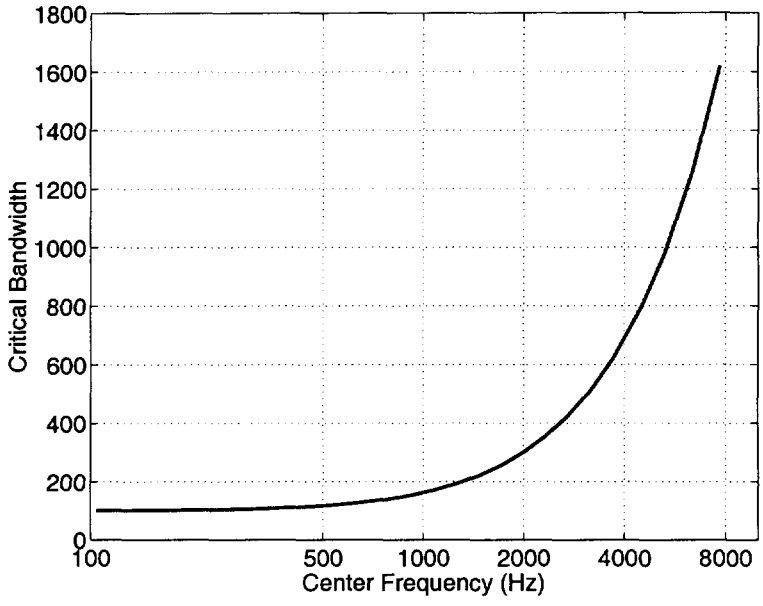


Figure 3.1: Variation of the critical band width as a function of the center frequency.

within a small frequency range centered around the frequency of the pure tone. In the case of high frequency tones, sounds within a relatively wide range of frequency around the frequency of the pure tone can mask the stimulus. From the above discussion, it is clear that auditory system uses non-uniform resolution in perceiving sounds. This resolution is high in the lower frequency region and low in high frequency region.

3.2.2 Simulating Critical Bands for ASR

It was observed in the previous section that the notion of critical bands suggests that the resolution of the human ear is non-uniform across the frequency band. As discussed in the previous chapter, this property of hearing is simulated in the current feature extraction module by projecting the power spectrum onto the frequency responses of the simulated

critical band filters. Figure 3.2 shows the basis functions onto which the power spectrum is projected to simulate critical band integration [14]. The set of filters illustrated in Figure

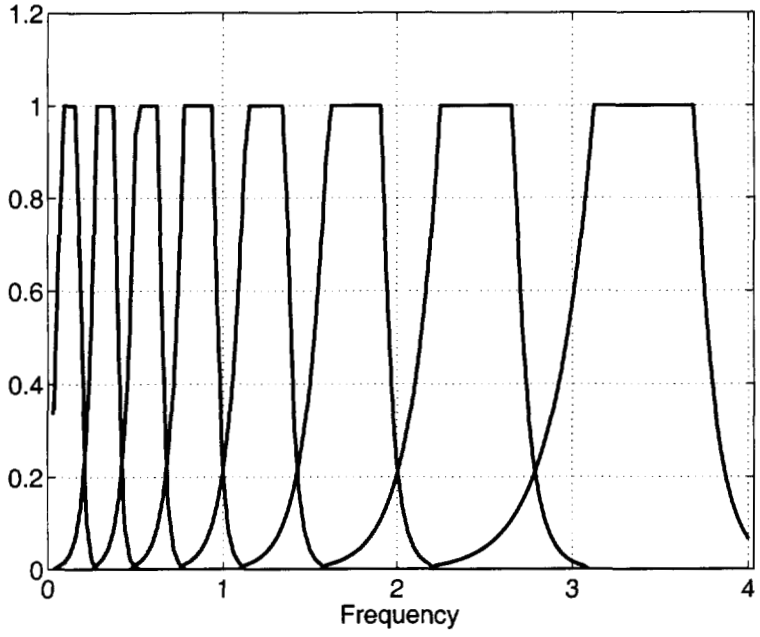


Figure 3.2: The basis functions used to simulate the critical bands. Note that only a few filters are shown to clearly illustrate the increasing band-width with centered frequency.

3.2 is used in perceptual linear prediction (PLP). The frequency response of these filters have a trapezoidal shape and their bandwidths are equal in the Bark frequency scale [14]. The Bark frequency Ω is related to frequency ω in the following manner [70].

$$\Omega(\omega) = 6 \ln \left[\frac{\omega}{1200\pi} + \left[\left(\frac{\omega}{1200\pi} \right)^2 + 1 \right]^{0.5} \right]. \tag{3.1}$$

Another widely used simulation uses triangular shaped filters and follows Mel frequency scale [71] instead of Bark. The Hz to Mel conversion is given by the following equation.

$$Mel(f) = 2595 \log_{10} \left(1 + \frac{f}{700} \right). \tag{3.2}$$

Thus both the Mel and Bark frequency scales expand the frequency axis in the lower frequency region and compresses the high frequency region. Filter-banks with equal bandwidths in these scales have the effect of providing lower resolution in the high frequency

region and higher resolution in the low frequency region. In the remaining part of this chapter the relationship between this non-uniform resolution and phonetic discriminability is investigated. The effect of such resolution in measuring distance between two spectra is also studied.

3.3 Analyzing the Nature of Discriminant Vectors

The nature of phonetic information carried by different parts of the spectrum needs to be analyzed in order to understand the relevance of Mel/Bark like warping. In this section we explain the use of linear discriminant analysis (LDA) to analyze the phonetic information carried by the different parts of the spectrum. For this we use the English portion of OGI STORIES corpus which consists of about 200 speakers, each speaking for about a minute. Short-time Fourier analysis on this database yields logarithmic spectra from over 600,000 frames of data[2]. First, only vowels were considered for discriminant analysis. Class conditional means and covariances were computed for 12 vowels. The mean spectral patterns corresponding to the 12 vowels are shown in Figure 3.3. The mean spectral patterns do not show the harmonic structure usually present in the spectrum of vowels. This is due to the extensive averaging done over multiple speakers and phonetic contexts. Also notice that, in general, the second and third formants are fused together, again caused by averaging. Out of the two spectral peaks exhibited by most of the mean patterns, the first one seems to be significantly sharper than the second.

The covariance of the mean vectors constitutes the between-class covariance C_b and the average of the class conditional covariances gives the within-class covariance C_w. As seen in the previous chapter, eigendecomposition of $C_w^{-1}C_b$ yields the discriminant vectors that point in the direction of maximum class separability.

Figure 3.4 shows these linear discriminant vectors. The first noticeable pattern in these basis functions is the higher emphasis given to the low frequency part (up to 2 kHz) of the spectrum. This suggests that this portion of the spectrum carries the majority of

[2] A 256 point DFT was used to compute the short-time Fourier spectrum. Since this analysis results in a symmetric spectrum, only the first 129 points were considered. The speech was sampled at 8 kHz and hence the 129 point spectrum gives a uniform resolution of 31.25 Hz.

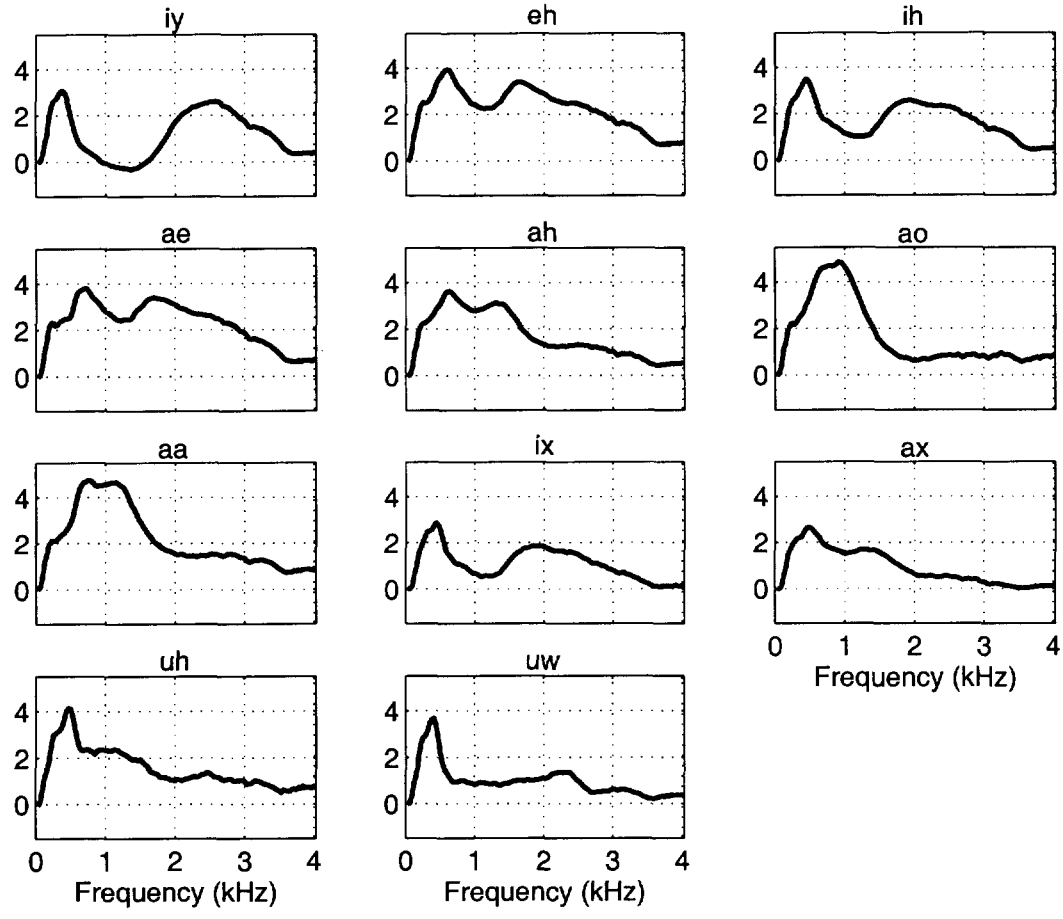


Figure 3.3: The mean spectral patterns computed for 12 vowels from the OGI STORIES corpus.

features relevant for discriminating vowels. Another important feature is that the higher order basis vectors have more cycles than the lower ones. Since the discriminant vectors are ordered with respect to the amount of discriminable information carried, it implies that the gross shape of the spectrum provides features that are more important than finer details. Within each discriminant vector the analysis resolution seems to drop with frequency, as indicated by the length of half cycles observed in the basis functions. For example, in the first discriminant vector, the periods of the first three half cycles are about 185 Hz, 375 Hz, and 1125 Hz, respectively. This clearly indicates an increased analysis resolution given to the lower frequency region as compared to the high frequency part of the

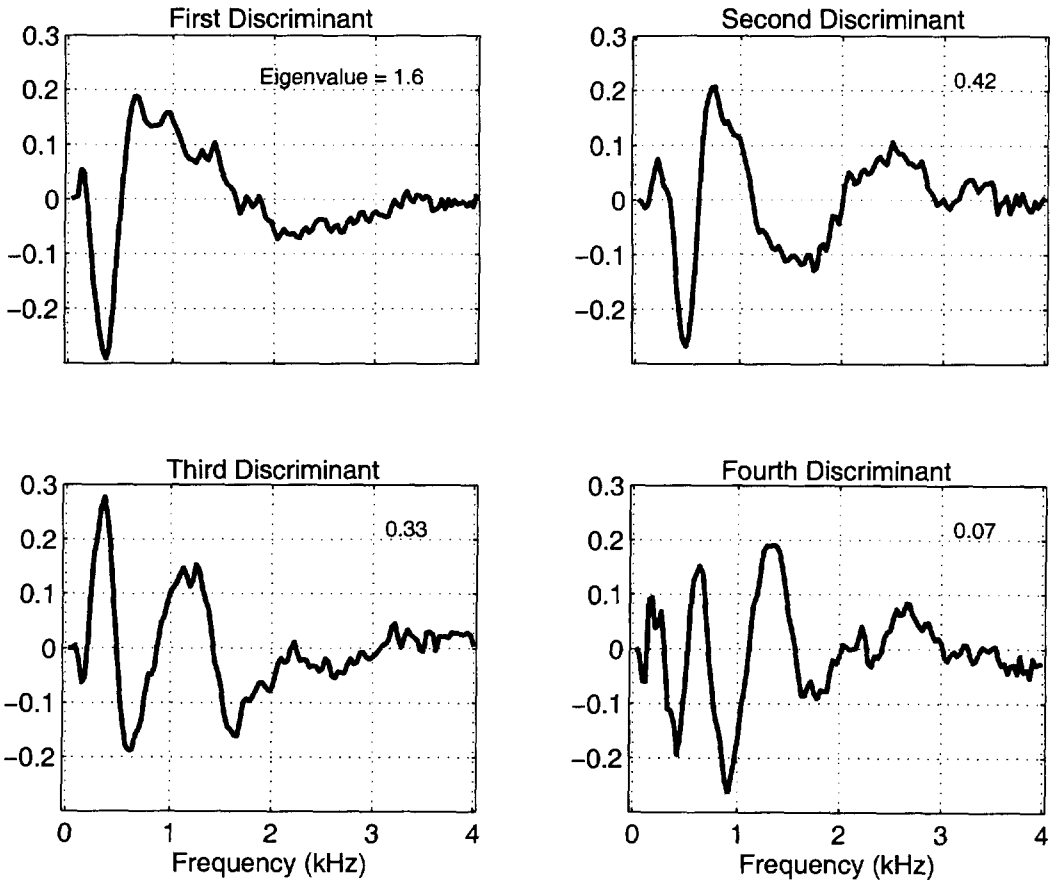


Figure 3.4: The first four discriminant vectors which describes directions in the feature space which provides maximum linear separability between vowels.

spectrum. Figure 3.5 shows the period of the half cycles, also known as the ripple period, as a function of frequency measured form the first six spectral discriminants. From the figure it is evident that the ripple period increases with frequency and hence the analysis resolution decreases with frequency. The trend is similar to the resolution provided by critical band analysis. This result indicates that the discriminative information in speech sounds is consistent with the non-uniform frequency resolution observed in human auditory system. Such non-uniform frequency resolution has also been suggested as a method to normalize vocal-tract length variations caused by speakers [72].

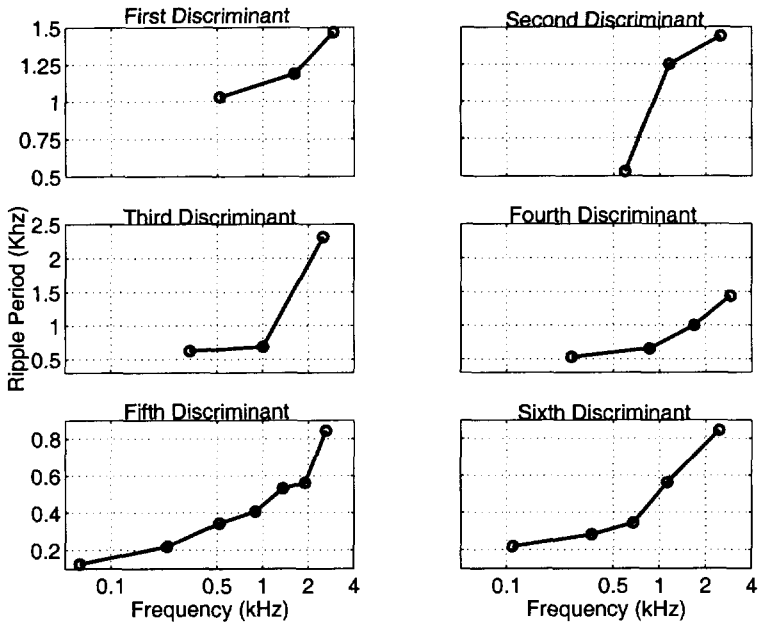


Figure 3.5: Plot showing the period of half-cycles (ripple period) as a function of frequency, measured from the first six discriminants.

3.4 Effect of Constituent Phonetic Classes on Discriminant Vectors

In this section, we describe the results of discriminant analysis on four different phonetic categories. The categories are vowels, stops, fricatives and diphthongs. The phonemes involved in these broad categories are listed in Appendex A. This experiment is designed to investigate the relationship between the nature of the discriminant vectors and the phonemes present in the broad categories.

Figure 3.6 shows the first four discriminant vectors derived separately from vowels, stops, fricatives, and diphthongs. Nasals were left out since there were only two classes present in the analysis (which would lead to only one linearly independent discriminant vector). From the figure it is clear that the non-uniform resolution is prominent only in the discriminant vectors extracted from vowel and diphthong categories. The discriminant vectors extracted from stops and fricatives seem to lack this behavior. This indicates that

the non-uniform resolution exhibited by the discriminant vectors indicates some characteristics which are specific to sonorant sounds. Although the discriminant basis functions

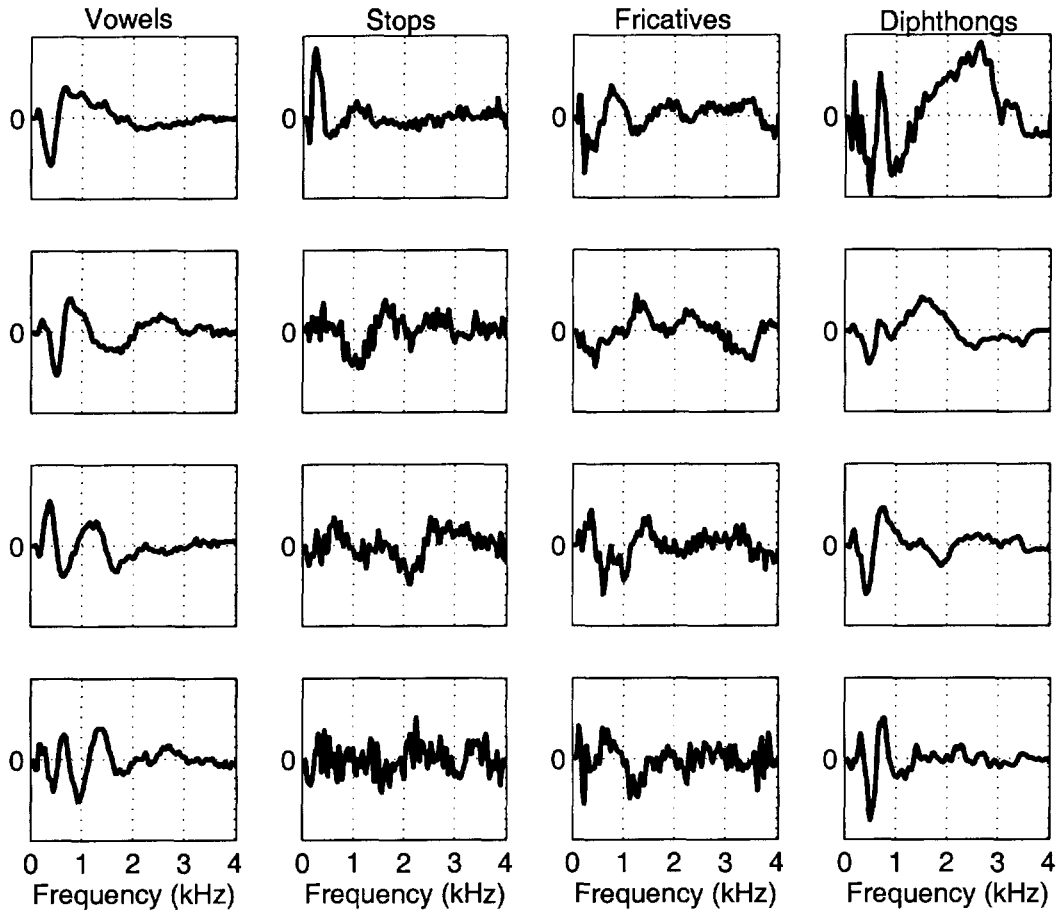


Figure 3.6: Linear discriminant vectors on sub-categories of phonemes. For vowels the discriminant vectors clearly indicate a higher analysis resolution in the low frequency region than in high frequency regions. This trend is absent for the stops and fricatives. In the case of diphthongs the non-uniform resolution is again evident especially in the first two discriminants.

derived from stops and fricatives are noisy the functions seem to exhibit certain interesting behavior. For example, the main difference between voiced and unvoiced stops is in the presence or absence of a spectral peak caused by voicing. This is illustrated in the mean spectral patterns of phonemes shown in Appendix B. The prominent peak in the first discriminant extracted from stop sounds could possibly detect the presence of voicing.

From an articulatory point of view, sonorants can be characterized by spectral features, where as the characteristic features of obstruents lie in the spectral dynamics. This is further supported by the F-ratio (ratio of across class to between class variance) given by the broad categories. Figure 3.7 shows the F-ratio given by the various broad categories. It is evident that vowels and diphthongs provide significantly higher F-ratios as compared to stops, fricatives, and nasals. This implies that there is limited benefit in using using the short-term spectrum as a feature to classify obstruent sounds.

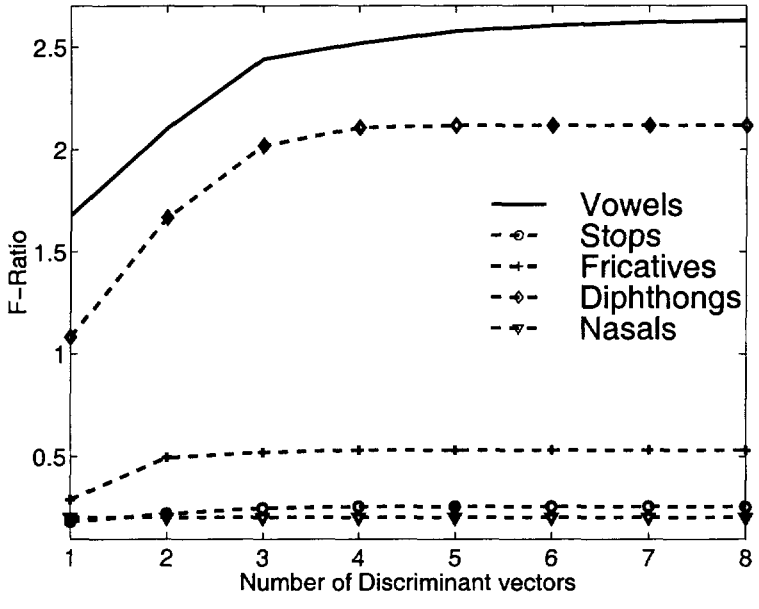


Figure 3.7: The F-ratio as a function of number of discriminant vectors used. The vowels and diphthongs possesses significantly higher linear-separability as compared to stops, fricatives and nasals.

3.5 Cause of Non-Uniform Resolution

Experiments in the previous section have demonstrated that the non-uniform resolution of the discriminant vectors is due to vowel-like sounds. Hence we analyze the functions that are optimal for discriminating two vowels. Figure 3.8 shows various pairs of mean spectral patterns and their corresponding discriminant functions. From Figure 3.8 we can see that

the spectral means of vowels are characterized by two spectral peaks[3]. It is also quite evident from the shape of the spectral means that the width of the spectral peak increases with frequency. For simplicity the class conditional covariances are assumed to be white. As shown in Appendix C, this implies that the the direction of maximum separability is given by the difference vector $d = \mu_1 - \mu_2$. From Figure 3.8 it is evident that the shape of these discriminant vectors clearly show non-uniform analysis resolution. It is shown in

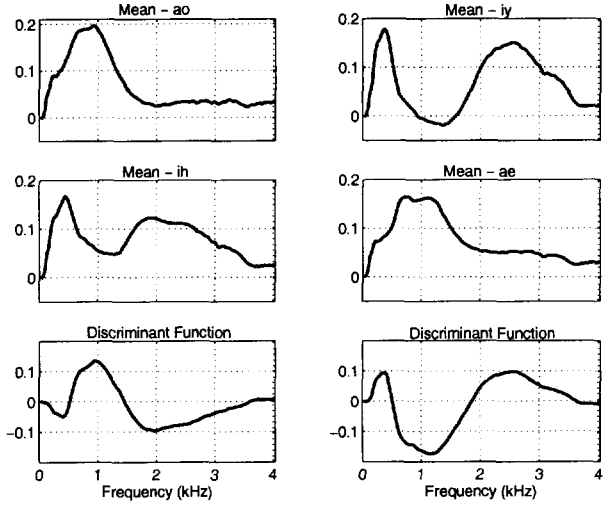


Figure 3.8: The class conditional spectral means of two pairs of vowels and the corresponding discriminant function.

Appendix D that this non-uniform resolution is caused by the broader peak that appears in the high frequency region of the spectral means.

Relation to Speech Production

In this subsection, the cause for the non-uniform width of peaks that appear in the spectral mean vectors is explained in terms of speech production. Note that these mean vectors are computed by averaging spectral vectors obtained from real speech data. Within a vowel class, the formant location varies from one spectral vector to another. One reason for this

[3]Note that the spectral peak and width of the peak are not directly related to the formant and bandwidth. The spectral shape corresponding to μ_1 and μ_2 has been averaged over many individual instances and hence need not physically resemble the spectrum of any speech sound.

variability is due to the differences in length of vocal tracts across multiple speakers. It can be shown that for a uniform tube model of the vocal tract, a change in the length of the vocal tract affects the higher formant locations more than the lower formants. Thus the variance of higher formant locations will be larger than that of the lower formants. Hence the peaks that appear in the high frequency region of the mean spectral vector will be broader. For a uniform tube model it can also be shown that the rate of change of logarithmic resonant frequencies with respect to the vocal tract length are equal for all the formants. This implies that for maximum linear discriminability, the frequency axis has to be logarithmically warped. The same effect can be achieved by using discriminant basis functions which provide non-uniform analysis resolution.

3.6 Effect of Non-Uniform Resolution

The cause of non-uniform frequency resolution exhibited by discriminant vectors has been established in the previous section. In this section, the effect of such basis functions is analyzed. Since most of the sonorant sounds are characterized by spectral peaks or formants, it is important to understand the sensitivity of these basis functions to formant locations. The logarithmic spectrum with a single peak (formant) with center frequency p and bandwidth b is given by the function,

$$s_p(f) = 10\log_{10}\left[\frac{1}{1+e^{(-b\pi T)^2} - 2e^{-b\pi T}\cos(2\pi T(f-p))}\right], \tag{3.3}$$

where T is the sampling period. A scalar feature, $y(p)$, extracted from the above spectrum using a basis function $d(f)$ is given by

$$y(p) = \int_0^{f_s/2} s_p(f)d(f)df. \tag{3.4}$$

The feature $y(p)$ is a function of the formant location p. Finally the sensitivity $s(p)$ of the discriminant function $d(f)$ is computed by

$$s(p) = \left|\frac{dy}{dp}y(p)\right|. \tag{3.5}$$

$s(p)$ gives the magnitude of the rate of change of the feature extracted as a function of the formant location p. Figure 3.9 shows the first two LDA basis functions and the

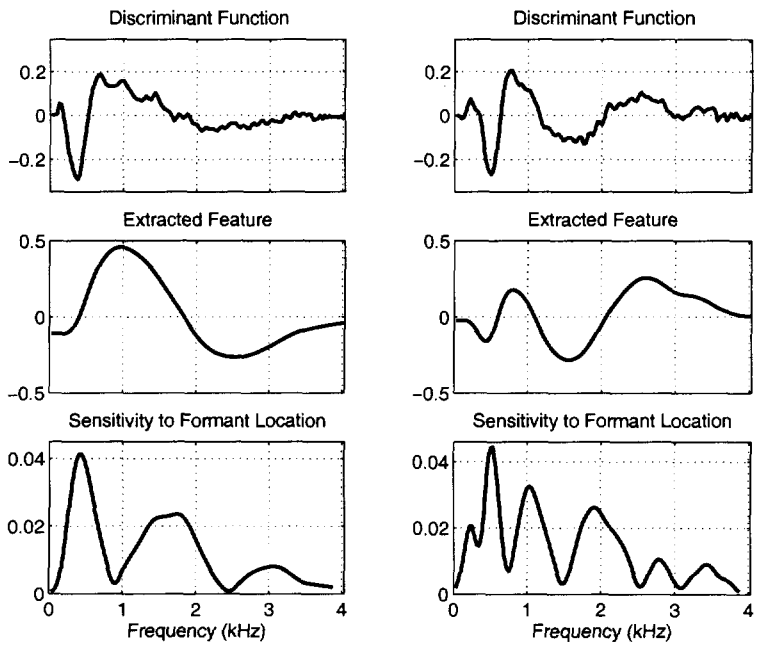


Figure 3.9: The sensitivity of two discriminant functions to the location of spectral peaks. The figure indicates a progressively decreasing sensitivity with frequency.

feature $y(p)$, extracted by them, as a function of the formant location p. The last row in the figure shows the sensitivity of these features to the formant location as defined by Equation 3.5. The sensitivity shows a cyclic behavior due to the sinusoidal nature of the basis functions. The amplitude of these cycles clearly show a decreasing trend with frequency. This demonstrates that the discriminant basis functions provide higher sensitivity to lower formants than to higher formants.

The above described sensitivity analysis was repeated with 8 dimensional features extracted by DCT and LDA and the PLP filter bank analysis. It is evident from the Figure 3.10 that features extracted by DCT provide relatively uniform sensitivity. The features extracted by LDA are more sensitive to changes in low frequency than in high frequency. Also note the the PLP filter bank analysis, that simulates the critical bands, also provides features which are more sensitive to changes in low frequency than in high frequency. Hence, the non-uniform sensitivity of LDA features seems to be consistent with human hearing.

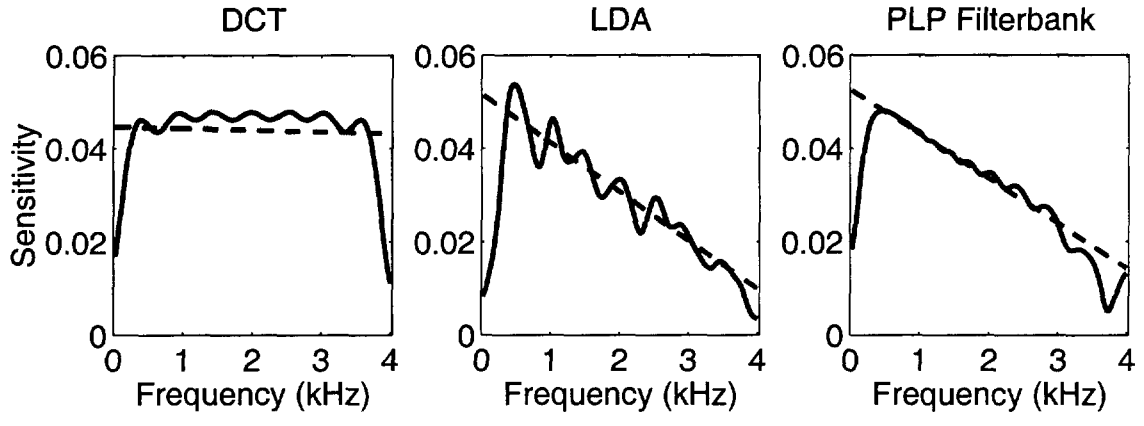


Figure 3.10: The sensitivity of feature vectors extracted by DCT, LDA and PLP filter bank analysis.

Non-uniform sensitivity of LDA features to formant location is consistent with the properties of human hearing. Using perceptual experiments Flanagan showed that the just noticeable difference in perceiving formants is proportional to the formant frequency [73]. This means that for humans to perceive a difference, formant shift has to be larger in terms of absolute frequency for higher formants than for lower formants. This illustrates that some of the properties of human hearing desirable for phonetic discrimination. This property has also been exploited by PLP modeling [14].

3.7 Connected Digit Recognition Experiments

In this section, the effect of discriminant basis functions on connected digit recognition is studied. These basis functions are directly applied on the DFT spectrum. Hence a comparison of its performance with the DCT will demonstrate the importance of the non-uniform analysis resolution exhibited by the discriminant functions.

For connected digit recognition, linear discriminant analysis was performed with the 22 phonemes necessary to describe the digits[4]. The resultant discriminant vectors are shown in Figure 3.11. Note that the first discriminant vector does a voiced-unvoiced detection.

[4]The discriminant basis functions were derived from the STORIES corpus.

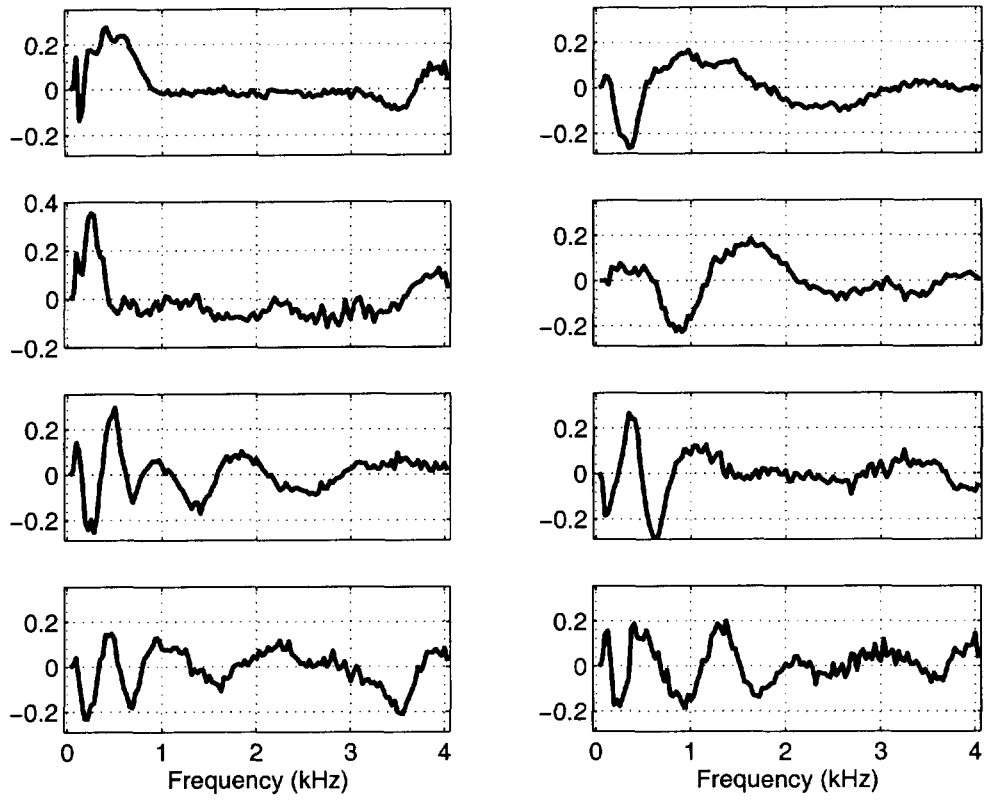


Figure 3.11: The first eight spectral discriminant functions derived from OGI STORIES data base. Only the 22 phonemes occurring in digits were considered for LDA.

The processing carried out by the first basis vectors to the portion of the spectrum that lies in the telephone bandwidth can be summarized as follows. Logarithmic energy between 300 and 900 Hz is integrated. Energy that falls above 3000 Hz is then subtracted from this quantity. Almost all the voiced sounds has more energy in the low frequency region than in the high frequency region. Hence the feature extracted by the first discriminant vector will be positive for voiced sounds and will be negative for unvoiced sounds. Hence we can conclude that the first basis vectors helps in making a voiced-unvoiced decision.

As seen earlier the higher discriminants provide more analysis resolution to the lower frequency region of the spectrum. These discriminant basis functions were used in a recognition experiment involving the OGI NUMBERS corpus. The training part consists of approximately 2500 utterances and the testing portion consists of 2200 utterances.

Table 3.1: Comparison of word error rates given by features extracted by projecting logarithmic spectra on cosine basis vectors and discriminant basis vectors. The task is to recognize continuously spoken digits from the OGI NUMBERS corpus. The table also shows the relative error reduction achieved by the use of LDA-derived basis vectors compared to cosine basis vectors.

HMM Structure	DCT	LDA	Relative Error Reduction
1-state 1-mixture	36.0 %	28.2 %	+21.6 %
3-state 3-mixture	11.1 %	7.9 %	+28.8 %
5-state 3-mixture	8.6 %	6.3 %	+26.7 %
5-state 5-mixture	7.8 %	6.1 %	+21.7 %

This database consists of different speakers speaking a string of continuous digits over different telephone channels. The number of continuous digits spoken varies from one to ten digits. The majority of these utterances are telephone numbers and zip-codes. For modeling, phonemes were considered as the basis acoustic units, and these were modeled using continuous density Hidden Markov models. The probability density for each state was modeled using mixture of Gaussian functions with diagonal covariances. The recognition experiments were repeated with various HMM architectures with varying number of parameters. The baseline system used the DCT on logarithmic power spectra. This was compared with a system that used the discriminant vectors shown in Figure 3.11 instead of cosine functions. Both the baseline system and the system using discriminant features used 8 dimensional feature vectors. In both cases, dynamic features consisting of velocity and acceleration coefficients were appended to the feature vectors. Table 3.1 shows the word error rates for the baseline system which used the DCT on Fourier spectrum and the system which uses discriminant basis functions. From the table it is evident that the discriminant basis functions significantly reduce the error as compared to the DCT. The relative reduction in error is always in excess of 20% irrespective of the complexity of the modeling used to represent the phonemes. For 3-state 3-mixture modeling the error reduction is almost 30%. An increase in model complexity leads to an overall reduction in error. For 5-state 5-mixture modeling the accuracy is above 92%. Even then, the discriminant basis functions provide a relative error reduction of 21%. This clearly indicates

Table 3.2: Word error rates for LDA features extracted from three different initial representations. Log spectrum refers to features obtained by projecting the logarithmic spectrum onto the discriminant basis functions. Discriminant features extracted after smoothing the power spectrum (either by PLP filter banks or by uniform filter banks) results in a minor improvement in performance.

Log spectrum	PLP filter banks	Uniform filter banks
6.3 %	5.8 %	5.8 %

the significance of Mel/Bark like resolution in achieving better recognition performance.

3.7.1 Significance of Smoothing the Power Spectrum

We have shown that the discriminant basis functions significantly reduce word error rate as compared to the DCT basis functions. In these experiments we have used logarithmic spectrum as the initial representation. As reported in the previous chapter discriminant analysis can also be performed on spectral vectors smoothed by critical-band filters. We observed that the error rate resulting from the LDA features derived directly from the logarithmic spectrum is higher than that resulting from the LDA features derived from the critical-band spectrum. This could be due to the integration done by critical-band filters on the power spectrum. In order to test this hypothesis we derived LDA features from filter banks with uniform bandwidth. Table 3.2 compares the performance of LDA features extracted from logarithmic power spectrum and from PLP filter banks. The table also shows the results for LDA features derived from uniform filter-banks.[5] LDA features extracted from uniform filter banks and from PLP filter banks performs equally well. This indicates the importance of integrating the power spectrum before passing it through the log nonlinearity.

The spectral integration that is performed before applying the logarithmic nonlinearity cannot be simulated by the LDA basis functions applied on the un-smoothed logarithmic spectrum since it requires nonlinear operations. Hence optimization of the filtering operation cannot be achieved using linear discriminant analysis. However, use of linear

[5]Forty uniform filter banks were used to derive the initial features. Linear discriminant analysis was performed on these features using the OGI stories data base as discussed in Section 2.

discriminant analysis makes the choice of these filters less critical.

3.8 Summary and Conclusions

In this chapter we investigated the nature of phonetic discriminant information present in the short-time spectrum of the speech signal. The shape of the discriminant vectors computed using LDA indicates the spectral variations that are important in discriminating phonemes. Analysis resolution of these discriminant vectors was determined by the period of a half cycle in the basis vectors. The smaller the period, the higher the resolution. We observed that the discriminant vectors analyze the low frequency part of the spectrum with higher resolution than the high frequency part. This trend is consistent with what is implied by using Mel/Bark filter banks. The cause of such non-uniform resolution as an outcome of LDA was further investigated. It was shown that this trend is primarily caused by vowels and diphthongs. Moreover, the resulting non-uniform resolution is attributed to the non-uniform width of peaks seen in the class conditional spectral means. The effect of such non-uniform resolution on the sensitivity of discriminant features to formant locations was also analyzed.

Non-uniform frequency resolution in the form of critical-band filters is mainly used in ASR systems to emulate human hearing. In this chapter we have established a link between non-uniform frequency resolution, phonetic discrimination and spectral properties of vowel-like sounds.

Chapter 4

Temporal Processing for Channel Normalization

4.1 Introduction

The focus of this chapter is on feature processing methods for increasing the robustness of speaker verification systems in the presence of channel variability. The variability introduced by microphones in particular and channels in general can significantly degrade the performance of both speech recognition and speaker verification systems. In automatic speech recognition(ASR) systems it is typically possible to train the recognizers with speech recorded using different telephone handsets and this makes the system relatively insensitive to handset variability. On the other hand in speaker recognition, a statistical model has to be designed from a relatively small amount of speech data to represent the acoustic features of the target speaker and it is not always practical to have training utterances collected from multiple telephone handsets. This results in models which are highly biased towards the handset used to record the training utterance. It has been observed that good verification performance can be achieved if the speaker uses the same handset and telephone line for testing [20]. On the other hand if the speaker uses a different telephone handset while testing, the verification error rate can increase by a factor of four to five [74, 22]. Hence features that are robust to handset variability are of significance in speaker verification.

In this chapter we introduce a data-driven method for designing handset normalizing

filters. The resultant filter is applied to the temporal trajectories of logarithmic spectral energies. Any stationary convolutive distortion will be an additive component in the logarithmic spectral energy domain, making it convenient for alleviating handset variability. Methods for processing temporal trajectories of logarithmic energy have already been proven to be effective in dealing with channel variability [23, 75]. In the current work we suggest a novel method to estimate the variability introduced by telephone handsets. This estimate is used to design a temporal filter which would suppress handset variability. Compared to some of the earlier methods [75], the proposed filter design method avoids the use of computationally expensive and time consuming speaker verification experiments.

The chapter is organized as follows. First a brief introduction to temporal processing is provided, followed by an analysis of the characteristics of handset variability. This leads to a discussion of the data-driven method used for the design of a handset *normalizing* filter in Section 4.4. Methods for the estimation of handset and phonetic variability are described in the following two sections, and these estimates are used by oriented principal component analysis (OPCA) in order to design an optimal filter. The characteristics of this filter are discussed in Section 4.6.

4.2 Introduction to Temporal Processing

This section provides a brief introduction to temporal processing techniques used to extract features from speech. For a detailed review refer to [27, 40]. Temporal processing in the context of this chapter means any modification to the temporal trajectory of a short-time spectral component. This spectral component is the output of an auditory-like (Mel or Bark) filter. Let us denote the spectrogram resulting from short-time analysis as $S(\omega_k, t_i), k = 1, 2 \ldots N, i = 1, 2 \ldots T$, where N and T are the number of frequency bands and the number of time steps used for the short-time analysis respectively. The temporal processing is performed on the logarithm of the squared magnitude, $S_l(\omega_k, t_i) = \log(|S(\omega_k, t_i)|^2)$. Then, $S_l(\omega_k, t_i), i = 1, 2 \ldots T$ is the time trajectory of the logarithmic energy corresponding to the k^{th} frequency band. The power spectrum of such a time trajectory is referred to as the *modulation spectrum* [27]. The Nyquist frequency

of the modulation spectrum is given by $\frac{0.5}{(t_i - t_{i-1})}$. Typically, $t_i - t_{i-1}$ (which is the window shift in the short-time analysis) is 10 ms. Hence the Nyquist frequency of the modulation spectrum is 50 Hz. Figure 4.1 illustrates the concept of the modulation spectrum.

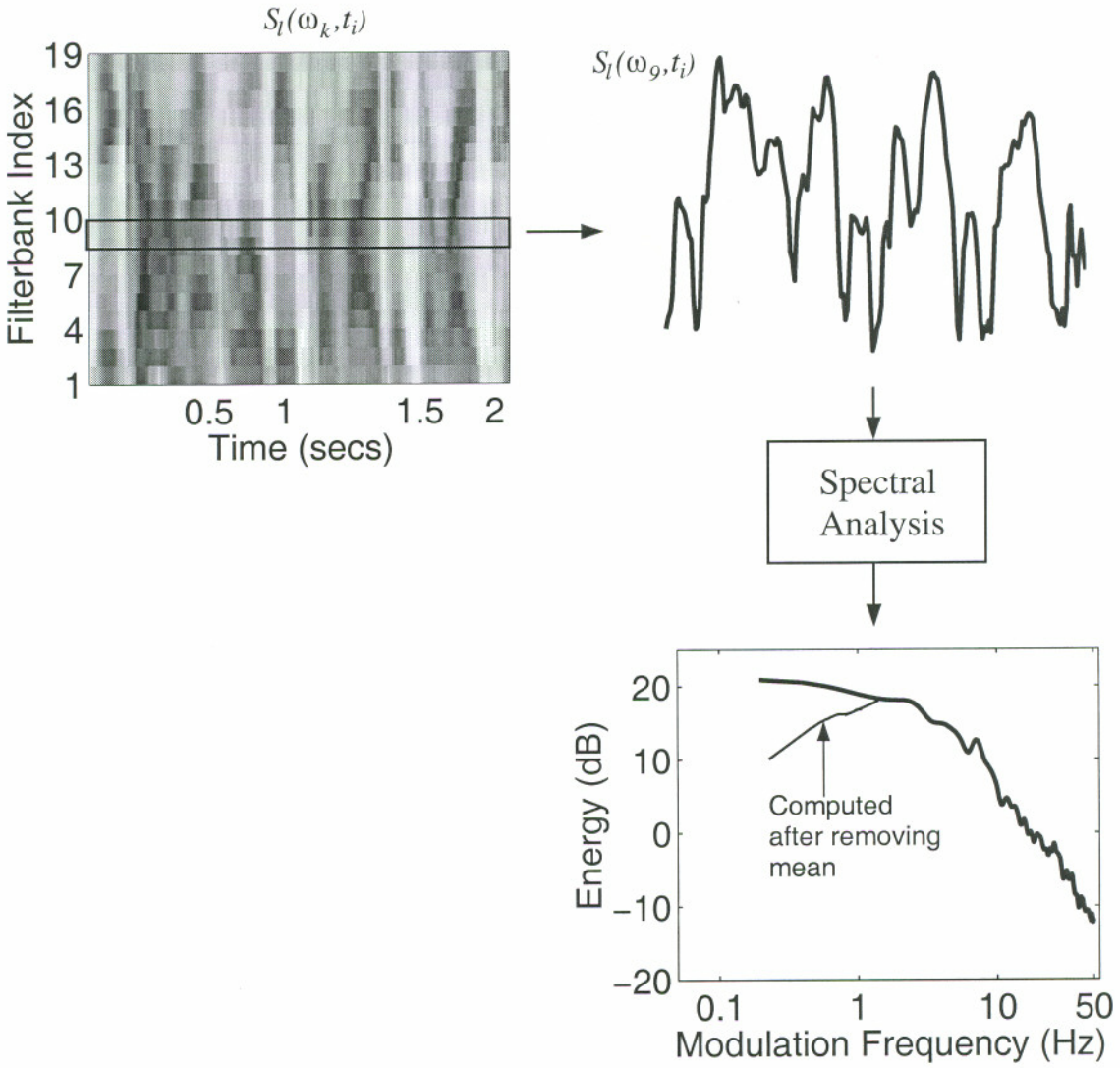


Figure 4.1: Illustration of the notion of modulation spectrum of speech.

If we assume that the telephone handset is linear, then the use of such a handset will introduce a convolutive distortion in the signal. This will show up as an additive component in the logarithmic spectrum of speech. A speech signal $s(t)$ recorded using a

microphone with impulse response $h(t)$ is transformed to $s_c(t)$ by

$$s_c(t) = s(t) * h(t). \tag{4.1}$$

After short term Fourier analysis, the short-term spectrum of the signal can be written as

$$S_c(\omega, t) = S(\omega, t)H(\omega) \tag{4.2}$$
$$\log(|S_c(\omega, t)|^2) = \log(|S(\omega, t)|^2) + \log(|H(\omega)|^2)$$

Note that we have assumed that the handset is stationary. Hence the short-term spectrum corresponding to the handset is independent of time t. $\log(|H(\omega)|^2)$ is the additive component introduced into the short-term logarithmic spectrum by the handset. This additive component could make the distribution of acoustic vectors significantly different from the distribution estimated from training data. This in turn leads to poor recognition performance. One way to deal with this is to use blind deconvolution [76] which removes the long-term average from the logarithmic spectrum.

It is straightforward to show that any convolutive distortion of the signal affects the mean of the time trajectory (direct current or the DC component of the modulation spectrum). Hence the mean subtraction (MS) operation applied on the log-magnitude spectrum (or any of its linearly transformed versions, like the cepstrum), widely known as cepstral mean subtraction (CMS) can be viewed as temporal processing which removes the DC component of the modulation spectrum. MS has been shown to make the features robust to handset variability [77, 3]. Mean removal can also be seen as a filtering operation where the filter is non causal and uses a variable number of past and future sample for the filtering operation. Figure 4.2 (a) shows the impulse and frequency response of the filter implied by mean removal. The frequency response is dependent on the length of the utterance. The high-pass cut off becomes lower as the length of the utterance increases. The MS has been proven to make the processing robust to handset variability in both speech recognition as well as speaker verification systems [77, 3], but the MS technique requires the entire utterance be available *a priori* and hence is not a suitable candidate for real time implementation.

4.2.1 Temporal Filters

More recently, RASTA (RelAtive SpecTrA) processing was introduced as an alternative to MS [23]. The RASTA filter is an IIR band-pass filter with the following transfer function

$$H(z) = 0.1z^4 \left[\frac{2z + z^{-1} - z^3 - 2z^{-4}}{1 - 0.98z^{-1}}\right]. \tag{4.3}$$

Figure 4.2 (b) shows the frequency response of the RASTA filter. The filter attenuates modulation frequency components below 1Hz and above 10 Hz. Thus, it not only eliminates stationary and slowly varying convolutive distortions, but it also eliminates fast varying (higher than 10 Hz) modulation frequency components. The low-pass filtering also helps to smooth spectral changes present in adjacent frames as a results of analysis artifacts (like position of the window with respect to the pitch period). RASTA-like filters have enjoyed considerable success in dealing with channel mismatches in ASR [23].

In the case of speaker verification, it has been shown that modulation frequency components as low as 0.125 Hz carry useful speaker information [1]. Hence use of the RASTA filter, which attenuates modulation frequency components below 1 Hz, degrades speaker verification performance. In order to address this problem, van Vuuren [1] uses a low-pass filter in conjunction with mean subtraction to normalize handset mismatches in speaker verification. Figure 4.2 (c) shows the impulse and frequency responses implied by such a technique. The design of this low-pass filter was based on performance of a speaker verification system. Hence the technique could be tuned to the characteristics of the recognition system and the data used to evaluate the performance. In the current work, we suggest a data-driven method to automatically design a temporal filter that suppresses handset variability. This method avoids the use of computationally-expensive and time-consuming speaker verification experiments [75] in order to design appropriate temporal filters. Since no explicit recognition is involved in the filter design, the result may also generalize to various recognizers.

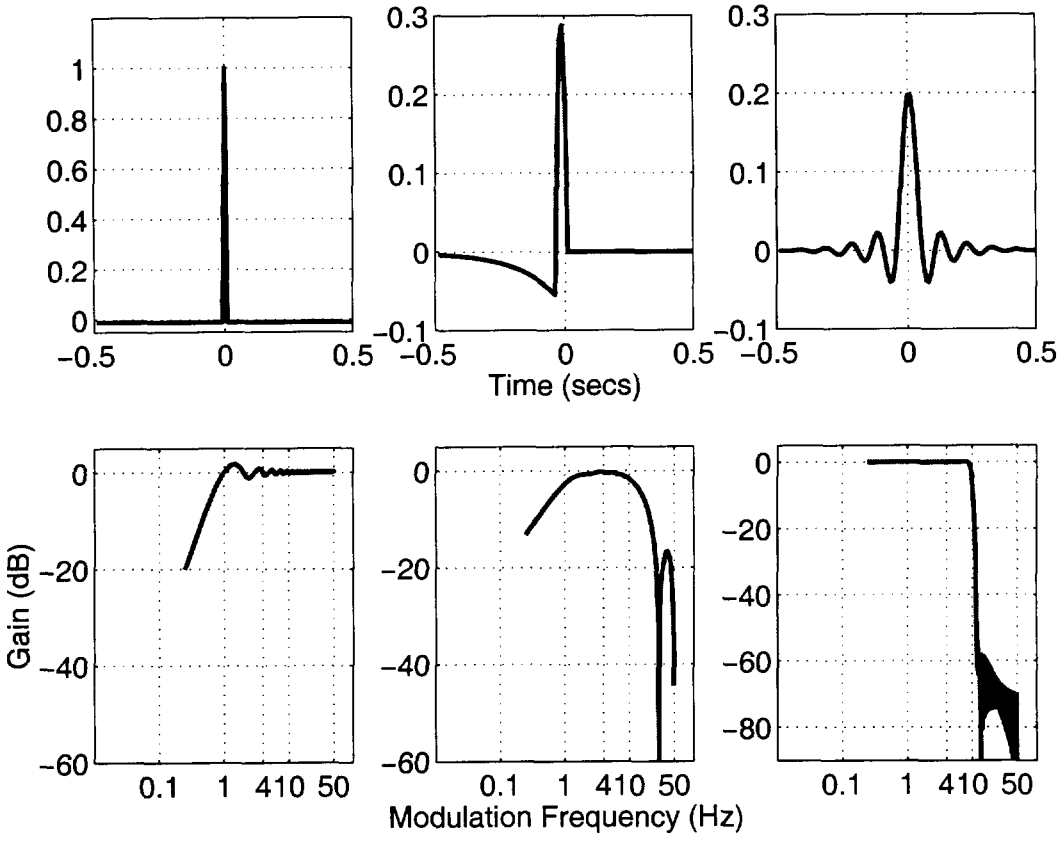


Figure 4.2: Impulse and frequency response of some existing channel normalizing temporal filters. (a) The filter implied by mean removal. The high pass cut-off frequency is dependent on the length of the signal on which mean removal is applied. (b) RASTA filter. (c) The temporal low-pass filter used by van Vuuren [1] in conjunction with mean removal for handset normalization in speaker verification.

4.3 Database to Study Channel Variability

The variability introduced by different telephone handsets significantly degrades the performance of speaker verification systems [74, 22, 78, 79]. Quatieri et al. studied the spectral differences introduced by carbon button and electret type microphones [80]. They found that carbon button microphones introduce what they call *phantom formants* due to the non-linear transfer function of the microphone. This demonstrates the severe mismatch that could result from changing the microphone type. Hence the study of handset variability reported in this chapter focuses on telephone handset variability.

For the analysis and estimation of handset variability, the HTIMIT database is used [81]. This corpus consists of speech utterances from 192 males and the same number of females recorded through various telephone handsets. It consists of 10 utterances per speaker. The HTIMIT corpus was collected by passing TIMIT [82] utterances through ten different microphones. Out of these ten microphones, four are of electret type and four are of carbon-button type.

As discussed in the previous section one of the factors that gets affected by variation in microphones is the long term average of logarithmic spectrum. Figure 4.3 shows the mean vectors of short-term logarithmic spectrum, computed from the same utterance passed through six different microphones. From the figure it is clear that the mean spectral vectors differ significantly from one microphone to another. As a whole, the spectral means corresponding to the carbon button microphones seem to be significantly different from that corresponding to electret type. Variability of the spectral means among carbon button microphones seems to be significantly higher than the variability among electret microphones. This is presumably due to the physical mechanism used in a carbon button transducer to convert acoustic pressure to electric signal.

Since in this work, temporal processing is used to normalize handset mismatches, the effect of different microphones on the temporal domain is also of interest. Figure 4.4 shows the temporal trajectory of logarithmic energy corresponding to the 5^{th} Mel filter bank. The major difference between the two handsets is the difference in the frequency-dependent gain which shows up as a DC shift in the logarithmic domain. It can also be

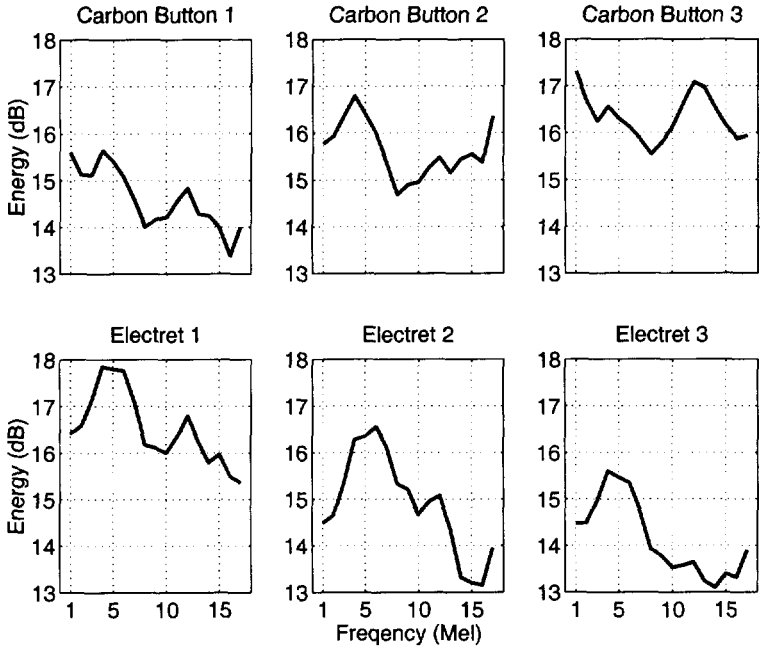


Figure 4.3: Spectral means extracted from an utterance recorded through six different microphones.

noted that this DC shift is non-uniform across different segments of the utterance.

From the above discussion, it clear that the variation in microphone type affects the temporal trajectory of logarithmic energy in a complex fashion. The following section develops a method to mathematically analyze these variations and to design filters to suppress this variability.

4.4 Channel Normalizing Filter Design

The objective is to design a filter which, by acting on the time trajectories of logarithmic energies, will minimize the variability introduced by different microphones. The initial short time processing will yield logarithmic energies from 19 Mel spaced filter banks, which are represented by $S_l(\omega_k, t_i), k = 1, 2 \ldots 19$. A one second long time trajectory corresponding to the kth Mel filter is denoted by,

$$\mathbf{x}_k(t) = \begin{bmatrix} S_l(\omega_k, t-50) & S_l(\omega_k, t-49) & \ldots & S_l(\omega_k, t+50) \end{bmatrix}^T,$$

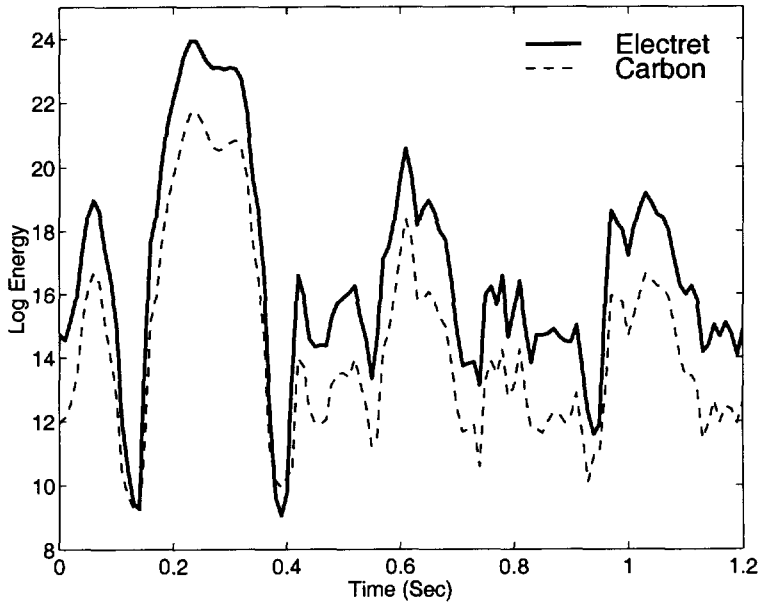


Figure 4.4: Temporal trajectories of logarithmic energies from the 5th Mel frequency band extracted from the same speech segment passed through carbon button and electret type microphones.

represents the signal that needs to be filtered. Since the window shift used in the short-term analysis is 10 ms, $\mathbf{x}_k(t)$ is a 101 dimensional vector.

The temporal filtering operation can be represented by the following equation.

$$y_k(t) = \mathbf{x}_k(t)^T \mathbf{h}_k$$

where $\mathbf{h}_k$ is the vector representing the time-reversed impulse response of the implied filter and $y_k(t)$ is the filtered signal. This filtering operation can be considered as projecting the vector $\mathbf{x}_k(t)$ onto the direction $\mathbf{h}_k$. The filtering has to be repeated for each of the the frequency bands. For simplicity, from here on we drop the index k representing the frequency band.

For handset normalization, $\mathbf{h}$ should point to the direction in the feature space where the variability due to handset is minimum. In order to prevent a trivial solution ($\mathbf{h} = 0$) we impose the additional constraint that $\mathbf{h}$ should retain as much of the relevant signal variability as possible, while suppressing the handset variability. The design criteria can

be achieved by selecting $\mathbf{h}$ to maximize the signal-to-noise ratio, ρ, given by the following equations.

$$\begin{aligned}\rho &= \frac{E[y_s(t)^2]}{E[y_n(t)^2]}, \\ &= \frac{E[(\mathbf{x}_s(t)^T\mathbf{h})^T\mathbf{x}_s(t)^T\mathbf{h}]}{E[(\mathbf{x}_n(t)^T\mathbf{h})^T\mathbf{x}_n(t)^T\mathbf{h}]}, \\ &= \frac{E[\mathbf{h}^T\mathbf{x}_s(t)\mathbf{x}_s(t)^T\mathbf{h}]}{E[\mathbf{h}^T\mathbf{x}_n(t)\mathbf{x}_n(t)^T\mathbf{h}]}, \\ &= \frac{\mathbf{h}^T\Sigma_s\mathbf{h}}{\mathbf{h}^T\Sigma_n\mathbf{h}},\end{aligned} \tag{4.4}$$

where $y_s(t)$ and $y_n(t)$ are the signal and noise components of $y(t)$ respectively. In the above equation Σ_s is the covariance of the signal and Σ_n is the covariance of the noise. Note that for this design we are considering the variance caused by handset as noise. It is fairly straight forward to show that the quantity ρ is maximized by setting $\mathbf{h}$ to $\mathbf{e}$, the leading eigenvector of the following generalized eigenvalue problem.

$$\Sigma_s\mathbf{e} = \lambda\Sigma_n\mathbf{e} \tag{4.5}$$

The above design process has to be repeated for each of the 19 Mel filter banks thus yielding 19 temporal filters ($\mathbf{e}_k, k = 1, 2 \ldots 19$), one corresponding to each of the 19 bands.

The generalized eigendecomposition implied by equation 4.5 is also referred to as oriented principal components analysis (OPCA) of the pair of random vectors, ($\mathbf{x}_s(t)$, $\mathbf{x}_n(t)$) [83]. It is so called since the principal eigenvector e is steered by the direction of $\mathbf{x}_n(t)$. It will be oriented towards the direction where $\mathbf{x}_n(t)$ has minimum variance while maximizing the variance of $\mathbf{x}_s(t)$. If the noise covariance is diagonal, $\Sigma_n = \sigma I$, then the noise vectors are isotropically distributed in the signal space. This will provide no steering to $\mathbf{e}$. In such a condition the generalized eigendecomposition $\Sigma_s\mathbf{e} = \lambda\Sigma_n\mathbf{e}$, reduces to the standard eigenvalue problem, $\Sigma_s\mathbf{e} = \lambda\mathbf{e}$. Hence it can be concluded that OPCA is not different from PCA unless the noise process $\mathbf{x}_n(t)$ is colored. Figure 4.5 illustrates this point. It is evident from the above discussion that the second order statistics of the handset and the desired signal variability are required in order to design handset normalizing filters.

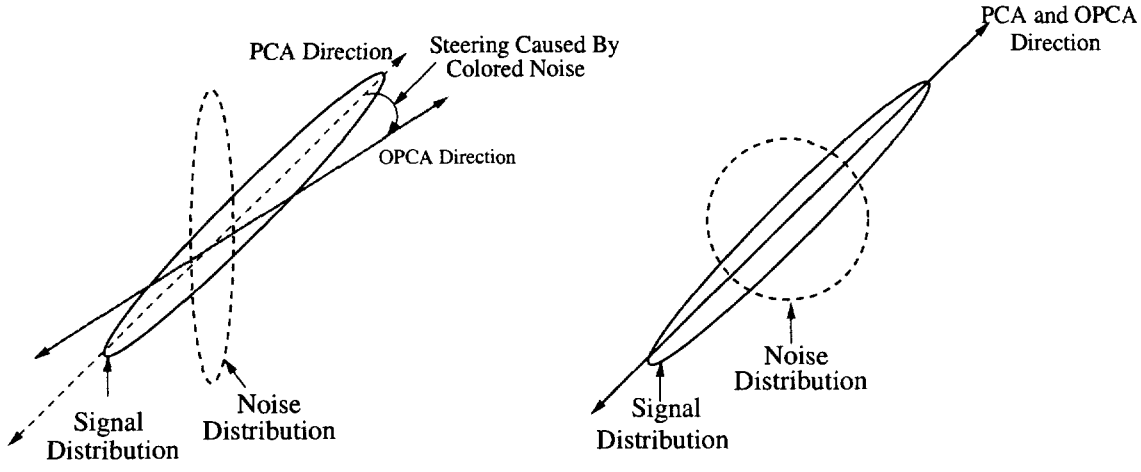


Figure 4.5: Illustration of the effect of colored noise in steering the OPCA away from the direction of noise variability. The figure also illustrates how OPCA defaults to PCA when the noise is white.

4.4.1 Estimation of Channel Variability

In this section a method to estimate the statistics of the handset characteristics is developed. The HTIMIT database is used for estimating the handset statistics. HTIMIT consists of TIMIT sentences passed through ten different microphones and they are recorded synchronously. For our experiments speech passed through four carbon button microphones and four electret microphones was used. Temporal feature vectors extracted from speech recorded through ith and the jth microphones are denoted by $\mathbf{x}_i(t)$ and $\mathbf{x}_j(t)$. It should be noted that $\mathbf{x}_i(t)$ and $\mathbf{x}_j(t)$ are extracted from the same sentence spoken by the same speaker in the same phonetic context. The only difference between $\mathbf{x}_i(t)$ and $\mathbf{x}_j(t)$ is that the speech signal from which they were extracted were recorded through two different telephone handsets. Hence the difference between these two temporal vectors d_k should point in the direction of microphone variability and is given by

$$\mathbf{d} = \mathbf{x}_i(t) - \mathbf{x}_j(t) \quad \forall i, j \quad i \neq j.$$

The noise vectors $\mathbf{d}$ that point in the direction of handset variability were computed over the entire HTIMIT database. This computation was independently performed on all the 19 Mel frequency banks. The following equation gives the covariance of the handset

variability, which is required for the design of handset normalizing filters.

$$\Sigma_n = E[\mathbf{d}\mathbf{d}^T] \tag{4.6}$$

4.4.2 Estimation of Desired Variability

For the design of the handset normalizing filter, we consider the variability introduced by the handset as the undesired source of variability. As discussed in the beginning of this section (Equation 4.4) in order to yield a non-trivial solution, the minimization of handset variability has to be carried out while preserving as much of the desired signal variance as possible.

Since our objective is text-independent speaker verification, it is tempting to consider the variability introduced by various speech sounds (which we call as phonetic variability) as undesirable. However, the simultaneous use of universal background model (UBM) and the speaker specific Gaussian mixture model (GMM) effectively alleviates phonetic variability [21]. Moreover, since each of the mixture components are independently adapted using speaker specific data, the GMM-based speaker verification system can potentially capture phoneme specific speaker characteristics. Hence removing phonetic variability could potentially degrade the performance. Therefore, in our filter design we are considering the variability introduced by phonetic classes as the signal variance that needs to be preserved[1]. For the filter to be optimal for speaker verification, we have to constrain the solution of Equation 4.4 so as to prevent the removal of speaker variability. This can be done by considering the combined speaker and phonetic variability as the desired source of variability. Even though the current design is not taking this aspect into account (hence in this respect is sub-optimal), the capability of the filter in making speaker verification robust to handset mismatches is experimentally demonstrated in the following section.

The phonetic variability that needs to be preserved in the 101 dimensional temporal vector $X_{kt}(i)$ is estimated as follows. Each of these temporal vectors is labeled by the phoneme that is aligned to the center element. The class conditional mean of the phoneme

[1]Even though the acoustic units defined by the UBM may not have an exact one to one correspondence with phonemes, we assume that preserving phonetic variability would lead to an improvement in the capability of the UBM in efficiently segmenting the acoustic space.

p is given by

$$\mu_p = \frac{1}{N_p} \sum_{\substack{t \in p \\ 1 \leq i < 8}} \mathbf{x}_i(t),$$

where the index i represents the handset index which ranges from 1 to 8 since we are using TIMIT sentences recorded using 8 different telephone handsets. If the edges of the temporal vectors span beyond the boundaries of an utterance then such vectors are ignored for the above computation. The phonetic variability corresponding to the k th handset is computed as the between-class covariance Σ_{sk} given by

$$\Sigma_s = \sum_p \frac{N_p}{N} [\mu_p - \mu][\mu_p - \mu]^T,$$

where N_p is the number of temporal vectors labeled as phoneme p, N is the total number of temporal vectors involved in the computation and μ is the global mean of the temporal vectors. The handset normalizing filter can now be estimated for each of the frequency bands using Equation 4.4. The next section discusses the characteristics of the estimated handset and phonetic variability.

4.5 Nature of Handset and Phoneme Variability

In the previous section we proposed methods for estimating handset and phonetic variability. Note that the OPCA approach to filter design is meaningful only if the handset variability (noise) is colored and is qualitatively different from the signal (phonetic). Hence in this section the nature of these two sources of variability are explored. They are analyzed only from the point of view of second order statistics (covariance) since we use only the second order statistics to design the filter.

Figure 4.6 shows the variance introduced by handset and phoneme into the temporal vector corresponding to the first Mel frequency band. The variance is nothing but the diagonal of the handset and phonetic covariance matrices, Σ_{nk} and Σ_{sk} respectively. From the figure it is evident that while the handset variability is constant across time, the phonetic variability has a peak at the center. As we move away from the center of the temporal vector, the variability reduces. This trend can be attributed to the procedure

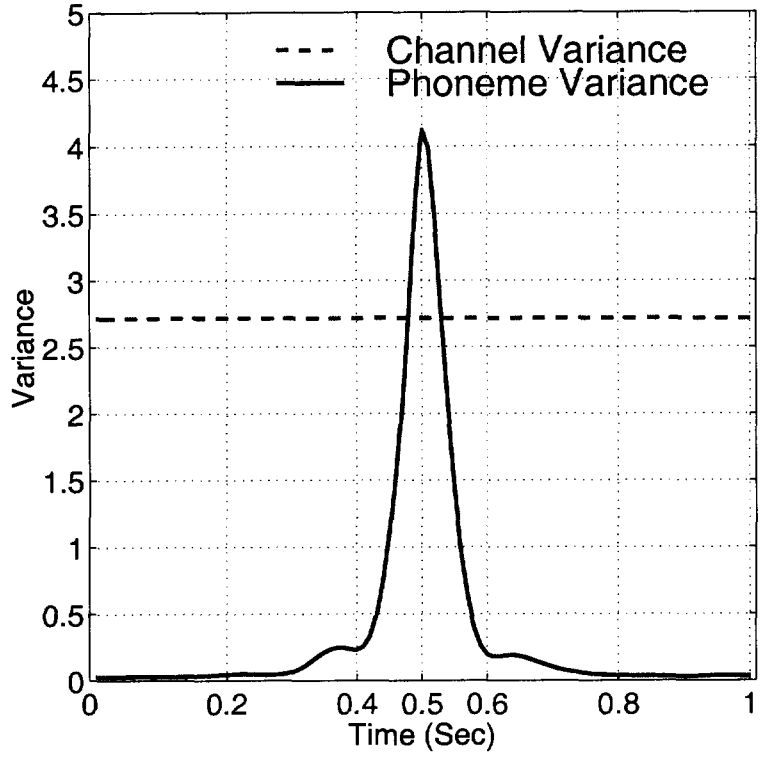


Figure 4.6: The contrasting nature of phonetic and handset variability.

used to compute the phonetic variability. Phoneme variability was defined as the variance of the class conditional means of the temporal vectors. A temporal vector is labeled to the class to which its center frame belongs. Hence the class conditional means of various phonemes show distinct phoneme-specific behavior at the center [84, 85]. Hence the variance of these means is higher in the center of the temporal vectors than at the edges.

In order to further investigate the nature of phonetic and handset variability principal component analysis was performed. Figure 4.7 shows the eigenvectors and the corresponding eigenvalues of the phonetic and handset covariance matrices. From Figure 4.7 the following conclusions can be drawn regarding handset and phonetic variability.

1. The Magnitude of handset variability is significantly higher than that of phonetic variability. This can be concluded from the higher eigenvalues associated with the

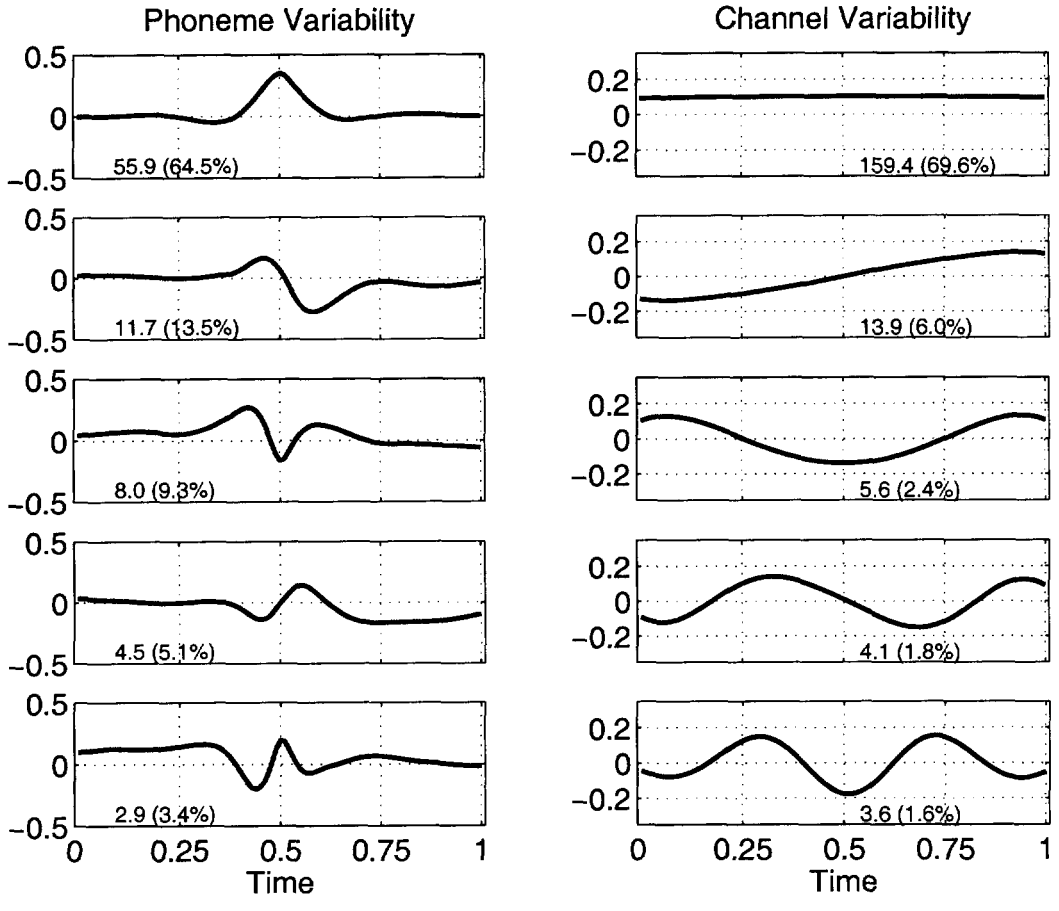


Figure 4.7: Eigenvectors corresponding to handset covariance and phonetic covariance. The numbers in each of the figures are the corresponding eigenvalues. The number in the bracket is the percentage of variability captured by the eigenvector.

eigenvectors of handset variability.

2. The leading eigenvector corresponding to handset variability is flat across time (DC) and it explains about 70% of the total handset variability. This means that the direction of maximum handset variability corresponds to the DC component of the modulation spectrum. Hence it can be concluded that 70% percent of handset variability is due to the DC component in the modulation spectrum. This means that methods like mean subtraction removes only about 70% of the variability introduced by handsets.

3. The directions of maximum phonetic variability are significantly different from that of handset variability. This justifies the use of OPCA in order to derive a filter to suppress the handset variability while preserving the phonetic variability.

4.6 Filter Characteristics

The 1^{st} eigenvector obtained from equation 4.4 is used as the handset normalizing filter. The filtering operation is carried out by projecting the temporal vectors, $\mathbf{x}_k(t)$, onto the eigenvector. The frequency response of the handset normalizing filters corresponding to the 19 Mel frequency bands are shown in Figure 4.8. From the frequency response, it

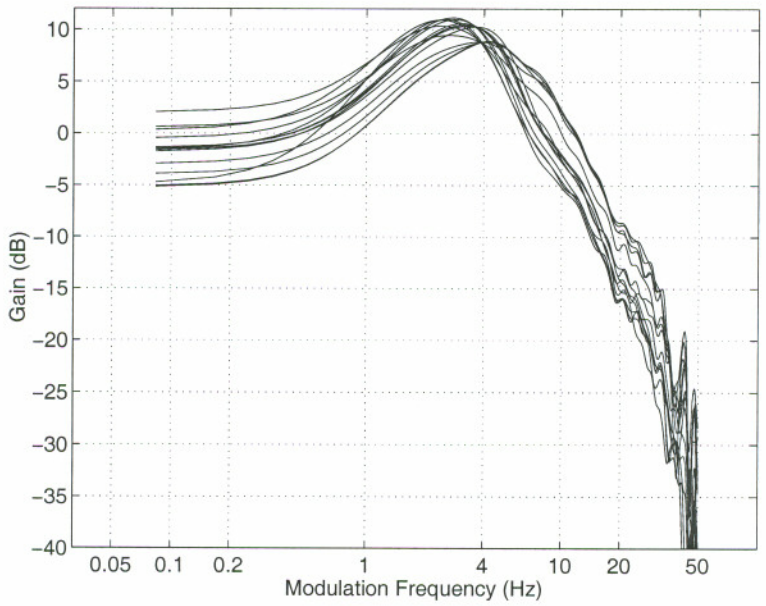


Figure 4.8: Frequency response of the data-driven handset normalizing filters

is clear that all these filters attenuate the DC component in the modulation spectrum. This property makes the filtered time trajectory insensitive to any linear convolutive distortion. Spectral components above 5 Hz are also attenuated. The filter emphasizes the components between 1 and 4 Hz. This is one of the main difference between this filter and the conventional RASTA filter, which passes the frequencies between 1 and 10 Hz (refer to Figure 4.2). The gain of the filter drops off at about 10dB/octave after 5 Hz. This

indicates that the frequency components between 1 and 4 Hz are robust for discriminating phonemes in the presence of handset mismatches. As these filters exhibit similar characteristics irrespective of the frequency band, for further discussion we will be considering only one of these filters (extracted from the 9^{th} band 900 Hz).

4.6.1 Comparison with LDA Filters for ASR

Earlier work by van Vuuren and Hermansky proposed a data-driven filter derived using LDA for improving the phonetic discriminability [86]. In their design, phonemes were considered as classes. The English portion of the OGI STORIES corpus was used to design these filters. One-second-long temporal feature vectors from each of 15 critical bands constituted the initial representation. Phonetic labeling information was used to classify each of these temporal vectors into one of the phonetic labels. Class-conditional means and covariances needed for LDA were computed for each of the phonetic classes, and this process was repeated for each of the 15 critical bands. The first three discriminant vectors given by LDA were then used as FIR filters and applied to the time trajectories of critical band energies. Out of the three filters the first (corresponding to the leading eigenvector) filter closely resembled the characteristics of the classical RASTA filter.

These filters were designed to preserve the phonetic variability while suppressing the within class variability. The within-class covariance contains the variability introduced by speakers, handset and the phonetic context (due to coarticulation) [62]. Hence the LDA filter suppresses not only the handset variability but also speaker and context variability, which causes the class conditional distributions of phonemes to overlap.

Figure 4.9 (a) shows the between-class and the within class variance. The between-class variance is significantly high for the central element of the temporal vector. The variance asymptotically becomes zero while moving away from the central element to the edge of the temporal vector. This trend is attributed to the procedure used to compute the phonetic variability. Phoneme variability was defined as the variance of the class conditional means of the temporal vectors. A temporal vector is labeled according to the class to which its center frame belongs, and the class conditional means of various phonemes show distinct phoneme specific behavior at the center [84, 85]. Hence the variance of these means are

higher in the center of the temporal vectors than at the edges.

Figure 4.9 (b) shows the handset+speaker variability and the variability introduced by phonetic context. Note that if these two are added the resultant quantity is equal to the within-class variability. From Figure 4.9 (c) it is clear that the temporal support of the discriminant filter is significantly shorter than that of the handset normalizing filter. Moreover the impulse response of the handset normalizing filter is smoother as compared to the discriminant filter. As a result of the relatively smaller temporal support, the pass-band of the LDA filter (1 Hz to 10 Hz) is much broader than the handset normalizing filter (1 Hz to 5 Hz). The shorter temporal support of the LDA filter is due to the design criteria which suppresses not only the handset and speaker variability but also the context variability. The context variability increases sharply while moving away from the center thus restricting the impulse response to a period of about 200 ms around the center of the impulse response. On the other hand, the handset normalizing filter is designed to suppress only the handset variability and hence is not influenced by the characteristics of the context variability. Hence the handset normalizing filter has significant activity for a period of about 500 ms making its frequency response more selective.

4.7 Effect of the Filter on Different Sources of Variability

Figure 4.10 illustrates the effect of the filter in normalizing handset mismatches. Figure 4.10 (a) shows the output from the 9^th Mel filter for a carbon button and an electret microphone. Figure 4.10 (b) shows the same after filtering. By comparing 4.10 (a) and 4.10 (b), it is evident that the mismatch caused by the use of different microphones has been significantly attenuated by the temporal filtering operation. Hence the data-driven filter is capable of suppressing variability introduced by handsets. It is also important to see the effect of filtering in the desired source of variability, which is variability introduced by phonemes. For this purpose phonetic and handset variability were estimated from the HTIMIT database after applying the handset normalizing filter. Figure 4.11 illustrates the effect of filtering on the handset and phonetic variability present in the HTIMIT database. Before applying the handset normalizing filter the handset variability is constant

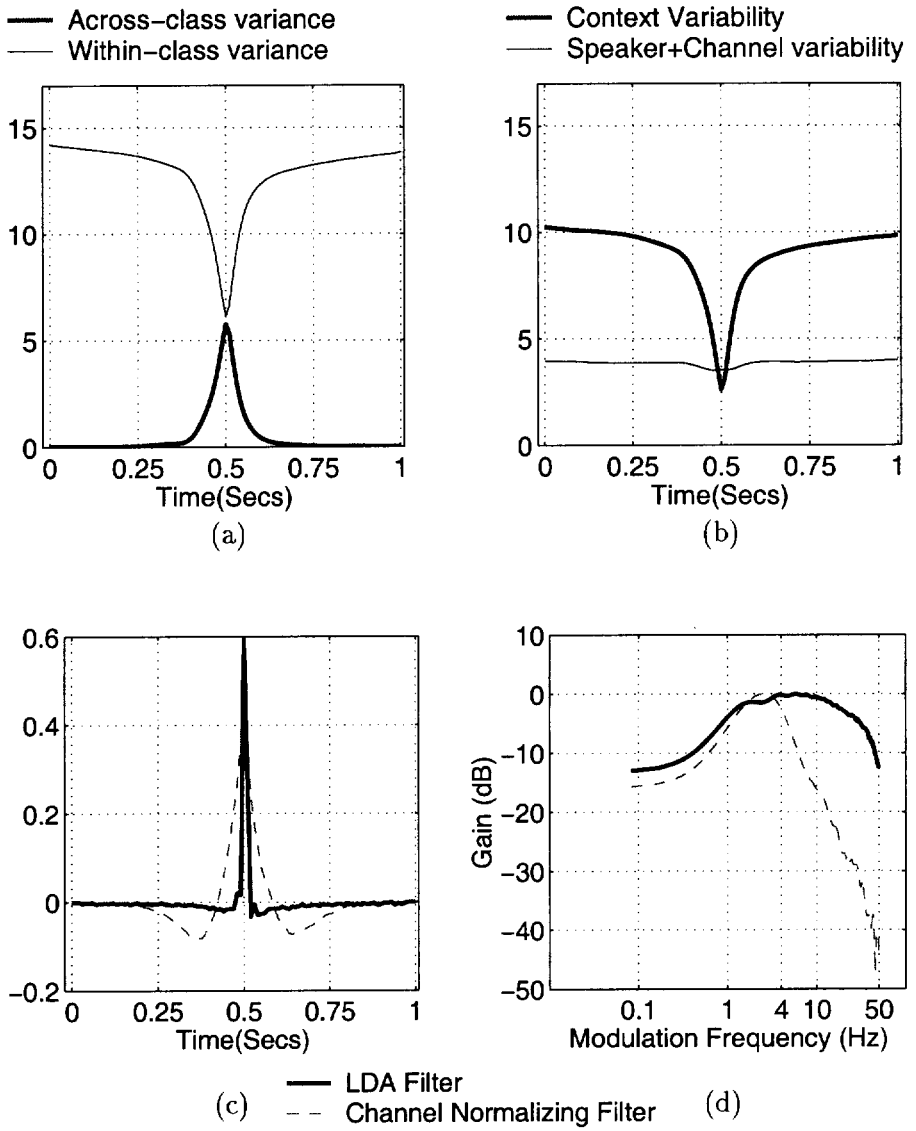
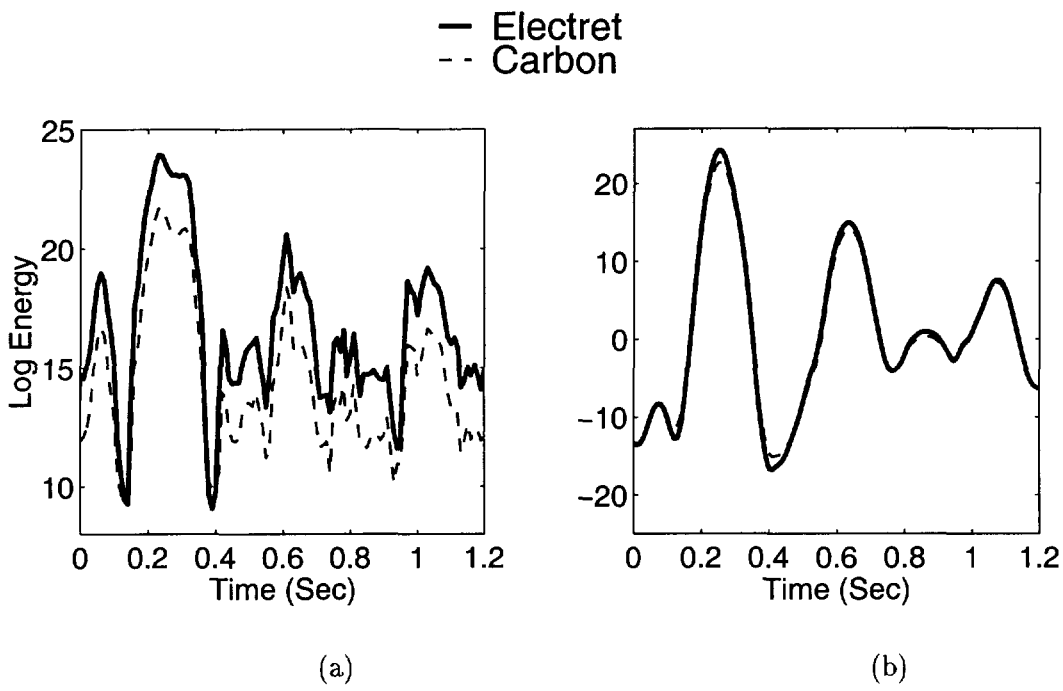


Figure 4.9: Comparison of the handset-normalizing filter with the temporal filter designed using LDA. (a) Within- and between-class covariances. Note that the between-class variance is maximum at the center of the temporal vector while it is minimum at the edges and vice versa for within-class variance. (b) Further decomposition of within-class variance into variances caused by handset+speaker and due to phonetic context. (c) Impulse response of the discriminant filter. (d) Frequency response of the discriminant filter.



(a) (b)

Figure 4.10: Demonstration of the handset normalization capability of the filter. (a) Time trajectories of the logarithmic energy before filtering. (b) The same time trajectories after filtering.

across time and it is significant as compared to the phonetic variability. On the other hand the phonetic variability is high at the center. After filtering, the magnitude of the handset variability is significantly reduced without any noticeable reduction in the phonetic variability. In fact, the handset variability is reduced to one tenth of its original magnitude. The filtering has also caused the phonetic variability to spread in time. This is due to characteristics of the impulse response of the filter. The significant reduction of variability introduced by handset without affecting the phonetic variability clearly shows the effectiveness of the proposed data-driven handset normalizing filter.

For an objective measure of the effectiveness, the signal-to-noise ratio was computed before and after the filtering. The signal-to-noise ratio ρ_k corresponding to the k^{th} frequency band after applying the filter f is defined by the following equation

$$\rho_k = 10 \log \left[\frac{f^T \Sigma_{sk} f}{f^T \Sigma_{nk} f} \right] \tag{4.7}$$

where Σ_{nk} is the noise covariance (handset variability) and Σ_{sk} is the signal covariance

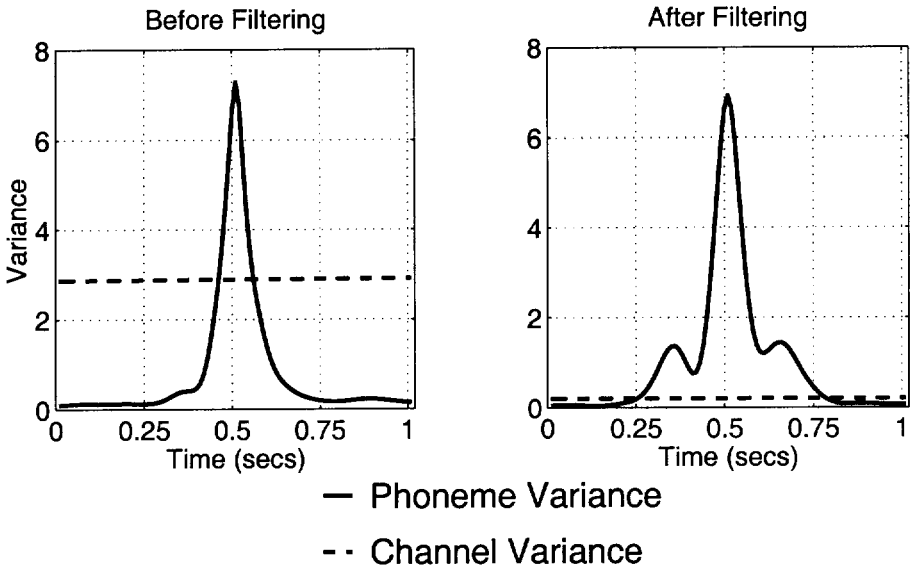


Figure 4.11: The effect of filtering on handset and phonetic variance.

(phonetic variability) extracted from the k^{th} frequency band. Figure 4.12 shows SNR as a function of the 19 Mel frequency bands before and after temporal processing. The RASTA filter significantly improves the SNR and the proposed data-driven filter improves the SNR even further. This illustrates the effectiveness of OPCA in designing handset-normalizing filters. It is also interesting to note that irrespective of the type of temporal filter, the SNR seems to be highest for the 5th and 6th Mel frequency bands. This is consistent with what was seen in Chapter 3 in the context of linear discriminant analysis of spectral vectors.

4.8 Summary and Conclusions

In this chapter we have developed a method for dealing with handset variability The handset normalization is achieved through filtering the time trajectories of logarithmic filter bank energies. A novel method to estimate the statistics of the variability introduced by the handset is also presented. The handset normalizing filter is designed to optimally suppress the handset variability under the constraint of preserving phonetic variability. The

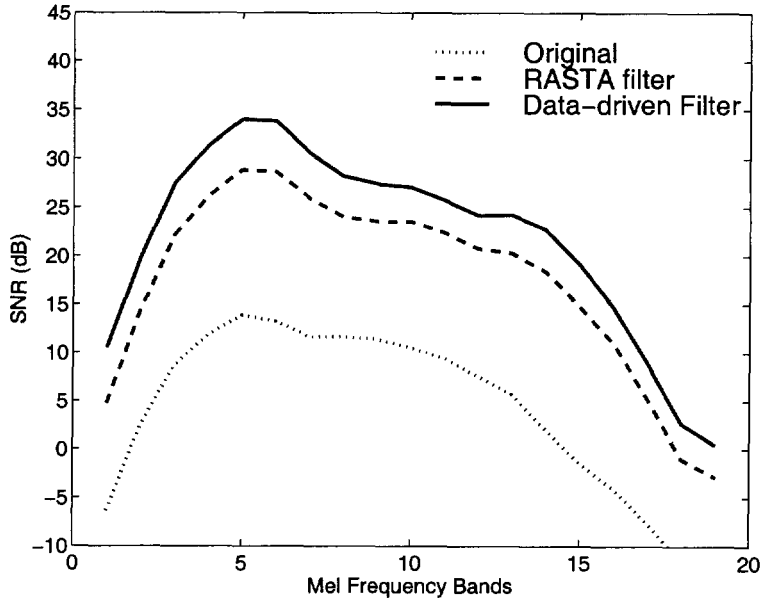


Figure 4.12: Signal-to-noise ratio as a function of the 19 Mel spaced filter-banks where signal and noise are defined to be phonetic variability and handset variability respectively. While RASTA filtering improves the SNR, the data-driven temporal filter makes the SNR even higher.

data-driven method uses only the second order statistics and thus the solution is obtained by solving a straight-forward eigenvalue problem. This filter attenuates the DC component in the modulation frequency thus making the features relatively insensitive to any stationary convolutive noise. The filter emphasizes the modulation frequency components between 1 and 4 Hz. The gain of the filter drops off fairly sharply beyond 5 Hz. We have shown that the filter is effective in suppressing the handset variability. The next chapter evaluates the capability of the filter in improving the robustness of speaker verification systems to telephone handset variability.

Chapter 5

Application of Handset Normalizing Filters in Speaker Verification

5.1 Introduction

In the previous chapter we discussed the design of handset normalizing filters capable of suppressing the variability introduced by telephone handsets. In this chapter we demonstrate the effectiveness of such filters in improving the performance of speaker verification systems. The performance of the data-driven filter is compared with conventional methods for channel normalization, like RASTA filtering, and with a recently proposed technique that uses low-pass filtering in conjunction with mean removal [1].

This chapter is organized as follows. Section 5.2 briefly covers the statistical modeling method used to model the feature vectors extracted from the speech of target speakers. Section 5.3 describes the details of the task and the database used for speaker verification experiments. This section also discusses the measure used to evaluate the performance of the verification systems. The effect of the handset normalizing filter in improving the speaker verification performance is discussed in detail in Section 5.4. This section also illustrates the effect of mean subtraction prior to filtering on both filter characteristics and speaker verification performance. In this section, the possibility of approximating the impulse response with mathematical functions is also explored. In Section 5.5 and 5.6, the significance of high and low modulation frequency components in the speaker verification performance is discussed. In the following section the performance of the proposed data-driven filter is compared with that of a low-pass filter specifically designed for improving

speaker verification performance.

5.2 Statistical Modeling for Speaker Verification

The task of speaker verification is to detect whether or not a given speech segment was spoken by the claimed target speaker. As mentioned in the first chapter, a speaker verification system can be text-dependent or text-independent. All the experiments described in this chapter address the text-independent speaker verification problem. In this section we briefly discuss the statistical modeling paradigm used for implementing the speaker verification system.

Systems that use a mixture of Gaussian functions for characterizing the distribution of acoustic vectors of target speakers have been shown to achieve good verification accuracy [20]. Hence for all our experiments the feature vectors extracted from the speech of the target speakers are modeled using Gaussian mixture models (GMM) given by

$$p(x_t|\lambda) = \sum_{i=1}^{M} p_i N_i(x_t; \mu_i, \Sigma_i), \tag{5.1}$$

where p_i are mixture weights, $N_i(\cdot)$ are Gaussian functions and M is the number of Gaussian functions used for modeling the distribution. The Gaussian functions are each parameterized by a mean vector, μ_i, and a covariance matrix, Σ_i. Diagonal covariances are used to speed up computation.

In text-independent speaker verification, the phonetic content of the utterance can vary significantly from the training utterance to the test utterance. Hence an effective method is necessary to suppress the variability introduced by the phonetic/linguistic content of the utterances. The unwanted linguistic variability of the acoustic features is suppressed by the simultaneous use of two GMMs, one is a speaker-specific GMM (λ_s), covering the acoustic space of a given speaker, the second is the so called universal background model (UBM) [21]. The universal background model (λ_b) is a speaker independent model, trained using the speech data of a large number of speakers. Thus the UBM represents a speaker-independent distribution of the feature vectors. The speaker-specific model is obtained by maximum a posteriori (MAP) adaptation of the UBM with the data of the

specific target speaker.

During the verification phase, the test utterance is scored using the UBM to obtain the acoustic transcription of the utterance in the form of a few top scoring mixture components. This transcription is used to restrict the scoring of the speaker dependent GMM to a small subset of mixture components [21]. The claim is rejected or accepted by comparing the log-likelihood ratio with a threshold θ as illustrated by the following equation.

$$\sum_i \ln \frac{p(x_i|\lambda_s)}{p(x_i|\lambda_b)} \underset{accept}{\overset{reject}{\lessgtr}} \theta, \tag{5.2}$$

where x_i is a feature vector extracted from the i^{th} frame. The relative tradeoff between false acceptance rate (FAR) and false rejection rate (FRR) is determined by the threshold θ.

5.3 Description of the Database

Continuous telephone quality speech from the SWITCHBOARD-2 phase 2 corpus was used for all the speaker verification experiments described in this chapter. This data set was used by National Institute of Standards and Technology (NIST) in its 1998 and 1999 official speaker verification evaluations [22]. The training data for each speaker consist of approximately two minutes of data collected over two one minute long sessions. The test set consists of utterances of durations 3, 10 and 30 seconds. For each of these testing durations results were analyzed for the following two different conditions:

- **Matched condition:** The training and testing utterances are collected from the same telephone handset. This condition is met only for the utterances of the genuine speakers (target trials). No such restrictions are placed on imposter trials.

- **Mismatched condition:** The test data is recorded through a handset which is of a different type compared to the one used for recording the training utterances. For example if the training utterance is recorded using a carbon button microphone then the test utterance of the target speaker is recorded using an electret type microphone. As with the matched condition, no restrictions are placed on imposter trials.

The errors made by a speaker verification system can be of two types, false acceptance (verifying a imposter speaker as the claimed target) and false rejection (rejecting a target speaker as an imposter). The trade-off between false acceptance rate (FAR or P_{FA}) and false rejection rate (FRR or P_{FR}) is determined by the threshold θ applied to the likelihood ratio score. A plot of FRR as a function of FAR is called the detection (DET) curve, illustrated in Figure 5.1. Equal error rate (EER) which is defined as the FAR (or FRR) when the FAR is equal to the FRR, is a widely used measure for evaluating the performance of the verification systems. NIST evaluates the performance of the systems using a measure called the decision cost function (DCF) where the cost of false alarms (C_{FR}) are 10 times more than the cost of false rejections (C_{FA}). The DCF is defined by

$$DCF = C_{FR}P_{FR}P_{Target} + C_{FA}P_{FA}P_{NonTarget}$$

where $C_{FR} = 10.0$, $C_{FA} = 1.0$, $P_{Target} = 0.01$ and $P_{NonTarget} = 0.99$. For more details refer to [87]. In our experiments, results were analyzed using both EER and DCF for the matched and mismatched test conditions.

5.4 Effect of Filtering

This section provides a detailed analysis of the effect of the handset normalizing filters on speaker verification performance. The objective is to investigate the effect of the proposed data-driven filter in improving the performance of the state-of-the-art speaker verification systems. Since the focus of the work is not on improving statistical modeling we use a standard GMM system [21] for modeling the speaker dependent distribution of the feature vectors. The task used for this is the official speaker recognition experiments suggested by NIST during the past two years [74, 22, 78].

5.4.1 Feature Extraction and Modeling

The feature extraction module used by the GMM-UBM based speaker verification has the general structure as illustrated by Figure 5.2. The initial processing involves the standard short-time Fourier analysis. Then the power spectrum is smoothed using simulated Mel

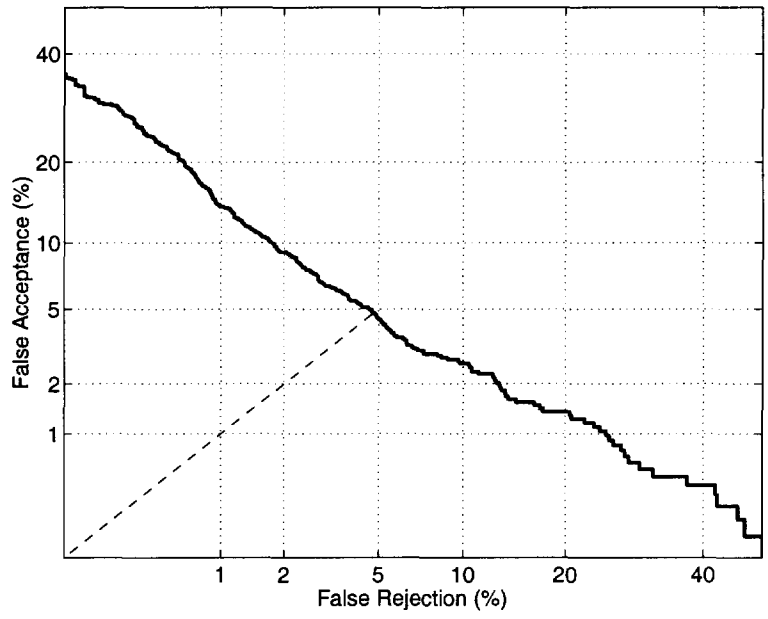


Figure 5.1: A typical detection curve (DET) curve. The point of intersection of the dotted line and the ROC curve corresponds to the operating point which yields equal error rate (EER).

filter banks. For the speaker verification experiments 19 filter banks falling within the telephone bandwidth were used. The data-driven filters are then applied on each of the time trajectories of logarithmic energies. From each of the 19 filtered time trajectories, delta features are computed [88]. The delta features are appended to the original feature vector to obtain a 38 dimensional vector. This is then whitened using a PCA transform computed from an independent data set. We have observed that whitening helps in reducing the number of mixture components without any degradation in performance.

For modeling the extracted feature we used the GMM-UBM paradigm which is described in [7]. A background model having 256 mixture components was trained using data from 80 speakers. This model was then adapted using the target speaker's data to obtain speaker-dependent models. For the implementation details of the system see [89].

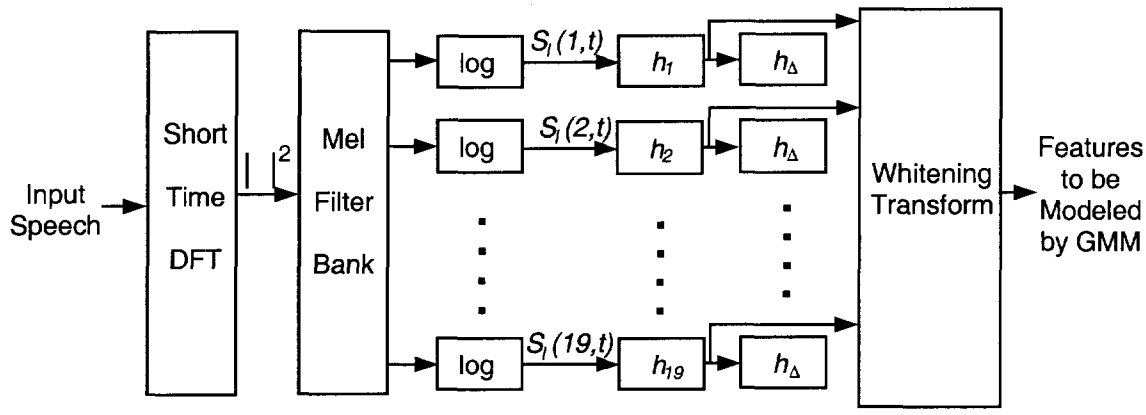


Figure 5.2: Block diagram of the feature extraction module used for speaker verification.

5.4.2 Experimental Results

Figure 5.3 illustrates the improvement brought about by the temporal processing. The baseline system does not use any filtering to deal with handset mismatches. RASTA processing improves the performance significantly in the mismatched condition. The EER is further reduced by the data-driven filter. As the duration of the test utterance increases, the performance of all the three systems improves, as expected. However the performance improvement in the case of baseline system is much less as compared to the systems that use temporal processing. Compared to the un-filtered case, the temporal processing helps in making the distribution of the training and the testing vectors more alike. Hence for the systems that use temporal filtering, more data directly leads to a reduction in error. In the case of the baseline system, even though the longer test utterances provide more data, its distribution still suffers a mismatch with that of training data. Thus longer test utterances are not able to reduce the error significantly. Due to this phenomenon, the relative error reduction achieved by the temporal processing increases with duration of the test utterance. For the 3 second condition, the data-driven filter yields a 19% reduction in error while the reduction in error becomes 35% for the 30 second test duration.

In the matched condition for 3 second long test utterances, both the filtering techniques seem to degrade the performance of the baseline system. This is attributed mainly two

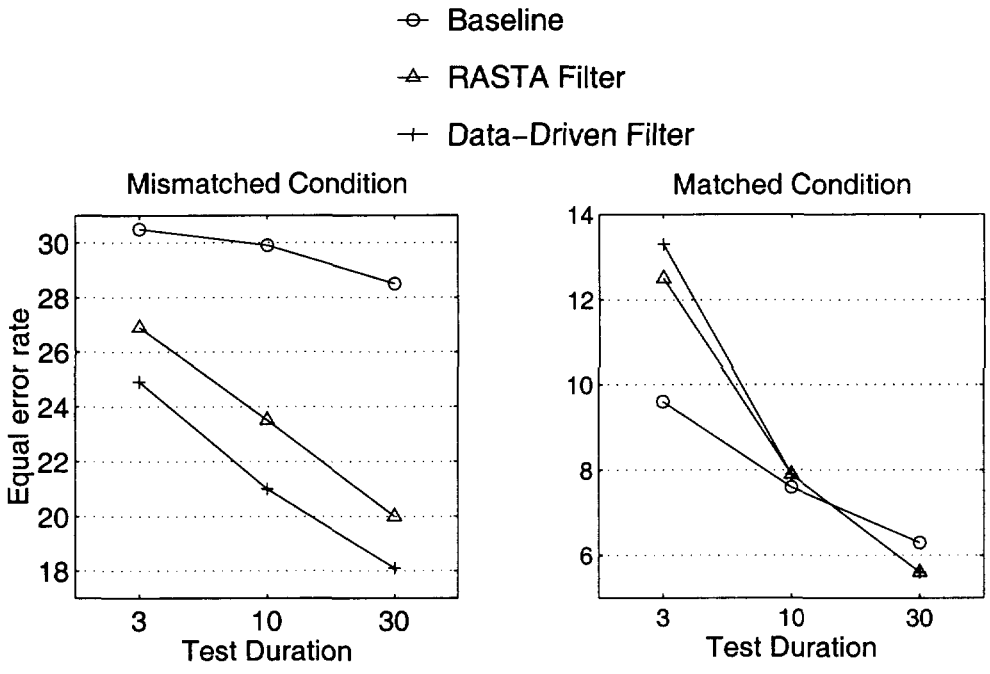


Figure 5.3: Plot illustrating the significant reduction in EER due to the use of temporal processing. The data-driven filter further reduces the EER as compared to the conventional RASTA filtering.

causes:

- In the matched condition, the verification system is capable of using the handset information in favor of target speakers to reject the imposters, who are guaranteed to use a different telephone line. Hence suppressing handset variability could lead to a degradation in performance in the matched conditions.

- Since the impulse response of the filter is 1 second long, the filtering artifacts (especially at edges) could be significant while filtering 3 second long utterances. This hypothesis is supported by the results, since the degradation caused by filtering decreases with increased duration of the test utterance. In fact for 30 second long test utterances, temporal filtering does provide a minor improvement over the baseline.

5.4.3 Approximating the Impulse Response using Difference of Gaussians (DOG)

By approximating the empirically derived impulse responses by a mathematical function we may reduce the noise introduced by the design processes. The choice of the function that can be used for the approximation is dependent on the shape of the impulse response. The impulse response of the filters used for the above experiments are shown in Figure 5.4. The impulse response of the filter resembles the shape of a difference of Gaussian (DOG)

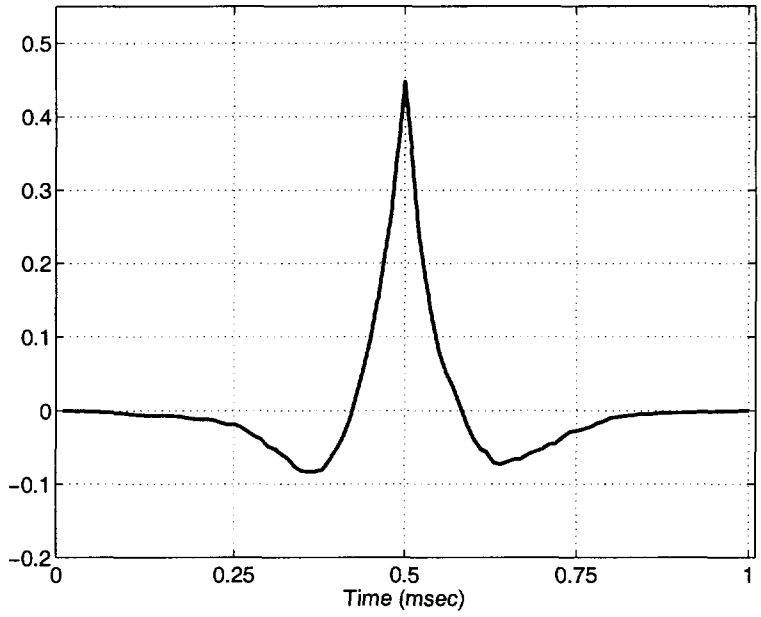


Figure 5.4: Impulse response of the handset normalizing filter. Shape of the filter is similar to of DOG (difference of Gaussians) filters used in image processing.

function. DOG function can be defined by two parameters, namely the two variances of the Gaussians,

$$D_g(\sigma_1, \sigma_2) = N(\mu, \sigma_1) - N(\mu, \sigma_2), \tag{5.3}$$

where $N(\cdot)$ represents a univariate Gaussian function and $D_g(\cdot)$ represents the DOG function, since the means of the two Gaussians are fixed. If the impulse response to be approximated by $G(.)$ is represented by h then we need to find σ_1 and σ_2 which minimizes

the squared error between $D_g(\cdot)$ and h. We define a mean squared error criteria given by

$$[\overline{\sigma_1}, \overline{\sigma_2}] = \text{argmin}_{\sigma_1,\sigma_2} \left[(D_g(\sigma_1, \sigma_2) - h)^T (D_g(\sigma_1, \sigma_2) - h) \right]. \tag{5.4}$$

The error minimization described by the above equation is carried out by the Gauss-Newton method [90]. The $\overline{\sigma_1}$ and $\overline{\sigma_2}$ that resulted in minimum least squared error were 21 and 173 respectively. The mean for the two Gaussians were maintained at the center of the impulse response (51 for a 101 point filter). Hence the handset normalizing filter h can be written as the difference of two Gaussians given by the following equation.

$$h(n) = \frac{1}{\sqrt{2\pi\sigma_1^2}} e^{-\frac{1}{2}\left[\frac{(n-\mu)}{\sigma_1}\right]^2} - \frac{1}{\sqrt{2\pi\sigma_2^2}} e^{-\frac{1}{2}\left[\frac{(n-\mu)}{\sigma_2}\right]^2}, \tag{5.5}$$

$$n = 1, 2 \ldots 101,$$

$$\mu = 51,$$

$$\sigma_1 = 21,$$

$$\sigma_2 = 173.$$

Figure 5.5 shows the result of approximating the impulse response in both frequency and time domains. The frequency response of the approximated filter matches well with the original frequency response in the regions where the gain of the filter is high. The DOG filter attenuates the high frequency much more than the original filter. The frequency components beyond 15Hz are attenuated to below -60 dB. This is due to the additional smoothing which resulted from approximating the filter using difference of Gaussians.

The DOG filter was used in place of the data-driven filter in speaker verification experiments. Table 5.1 compares the EER obtained by using the data-driven filter and the DOG filter. From the table it is clear that both the data-driven filter and its DOG approximation yield approximately the same performance. The DOG approximation seems to perform slightly better than the original filter in matched condition. This illustrates the feasibility of approximating the impulse response with DOG functions without affecting the performance.

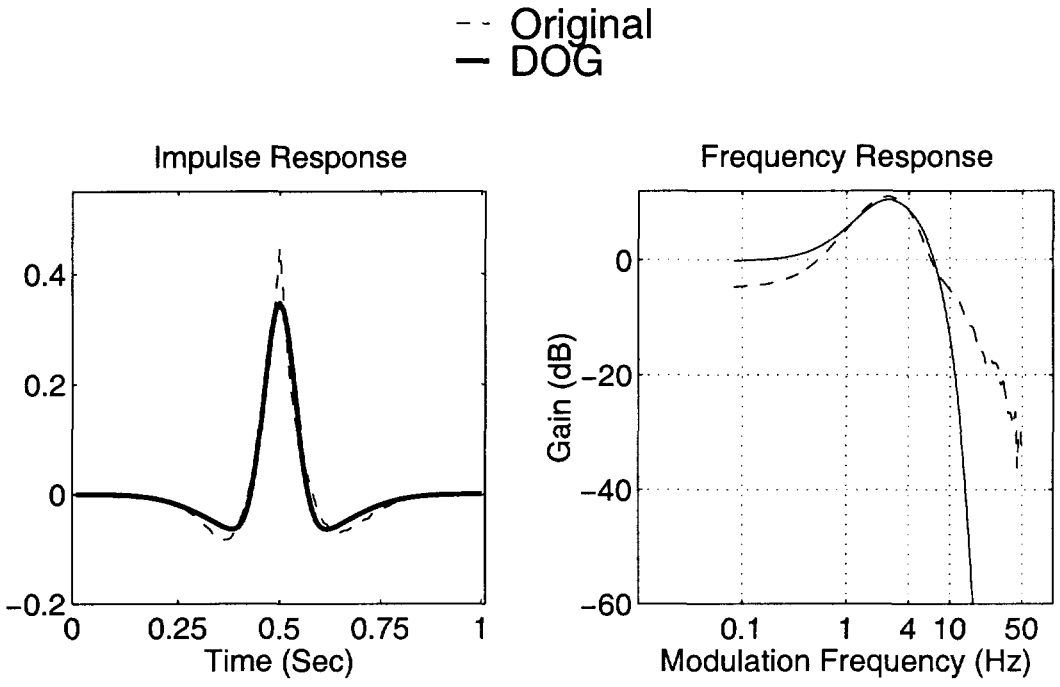


Figure 5.5: Comparison of the characteristics of the handset normalizing filter and the filter obtained by approximating the impulse response with DOG function.

5.4.4 Effect of Filtering in Addition to Mean Removal

In the past it has been shown that mean subtraction (MS) is an effective technique in dealing with channel mismatches [91, 77]. Unlike RASTA (which being a filtering technique can work on the signal online with a fixed delay) the MS technique needs the entire speech utterance prior to the processing. Thus the MS technique can be used in applications where there is the luxury of recording the speech data prior to processing. Nevertheless it would be interesting to see whether the proposed data-driven filters can be used to any advantage by applying them on mean subtracted time trajectories. For this purpose the handset normalizing filters were designed after applying an utterance based mean subtraction on HTIMIT. Figure 5.6 compares this filter to the filter derived without mean removal. The only major difference between the two filters is in the frequency characteristics below 1 Hz. The filter designed using mean removed data does not attenuate low frequency components. This suggests the lack of handset variability present in the modulation

Table 5.1: Comparison of data-driven handset normalizing filter and the corresponding DOG approximation in speaker verification performance.

Testing condition		Original filter	DOG filter
3 Sec	Matched	13.3 %	11.9 %
	Mismatched	24.9 %	25.2 %
10 Sec	Matched	7.9 %	7.5 %
	Mismatched	21.0 %	21.9 %
30 Sec	Matched	5.6 %	5.6 %
	Mismatched	18.1 %	18.4 %

frequencies between 0 and 1 Hz.

Figure 5.7 compares the performance of a system that uses only MS with a system which uses temporal filtering in addition to MS. It is evident that in the mismatched condition, the data-driven filter significantly improves the performance of the baseline system which uses only MS. This improvement becomes more significant as the test utterance duration increases. In the case of the matched condition, there is degradation caused by the filtering applied on top of MS. This degradation decreases as the duration of the test utterance increases. This clearly demonstrates that the data-driven temporal filter developed in this paper can yield significant additional improvements while using standard channel normalization techniques like MS.

5.5 Significance of High Modulation Frequency Components

In this section we analyze the performance of the data-driven filter used in conjunction with MS to demonstrate the significance of removing higher modulation frequency components for robustness. In order to assess how well the data-driven filters work on a new task, we consider the 1999 official NIST speaker verification task. The training data for each speaker consists of about 2 minutes of data collected from two separate (about 1 minute long) sessions. Testing consists of utterances of durations anywhere between 1 second to 1 minute. Figure 5.8 illustrates the improvement due to the data-driven temporal filter over the systems that use mean subtraction and RASTA filtering. In the mismatched condition RASTA filtering performs as well as MS. The data-driven filter applied in addition to

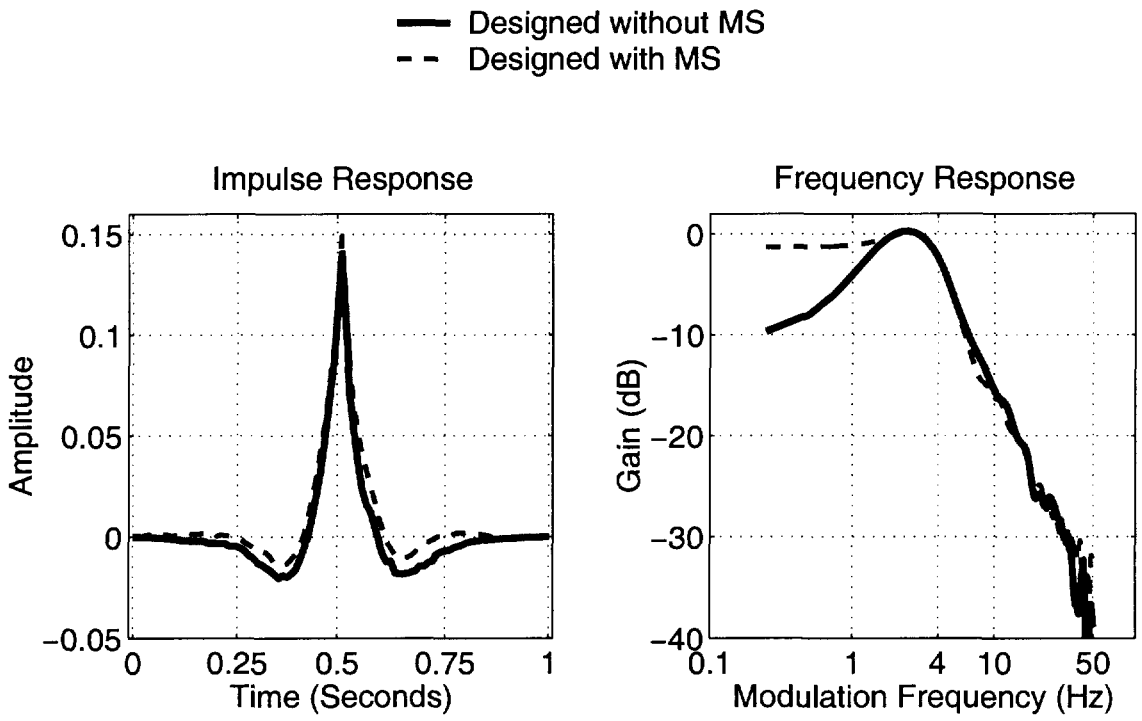


Figure 5.6: Effect of mean removal on the frequency response of the data-driven filter.

utterance-based mean subtraction yields a significant reduction in error when there is a handset type mismatch. This clearly illustrates the effectiveness of the proposed filter in suppressing the handset variability. In the matched condition, all of the three techniques perform equally well. Table 5.2 compares the EER and minimum DCF resulting from the use of MS, RASTA, and the data-driven filtering in combination with MS. In the mismatched condition, the data-driven filter reduces the equal error rate by more than 25 %. The MDCF is also reduced by about 10% as compared to RASTA. In the matched condition the data-driven filter performs as well as MS but RASTA causes a degradation in performance. This degradation is attributed to the removal of modulation frequency components between 0 and 1 Hz which have been shown to contain important information relevant for speaker verification [75]. Notice that while MS passes almost all the frequency components other than DC, the RASTA filter passes frequency components between 1 and 10 Hz. Compared to MS and RASTA filtering the data-driven filter has a much lower

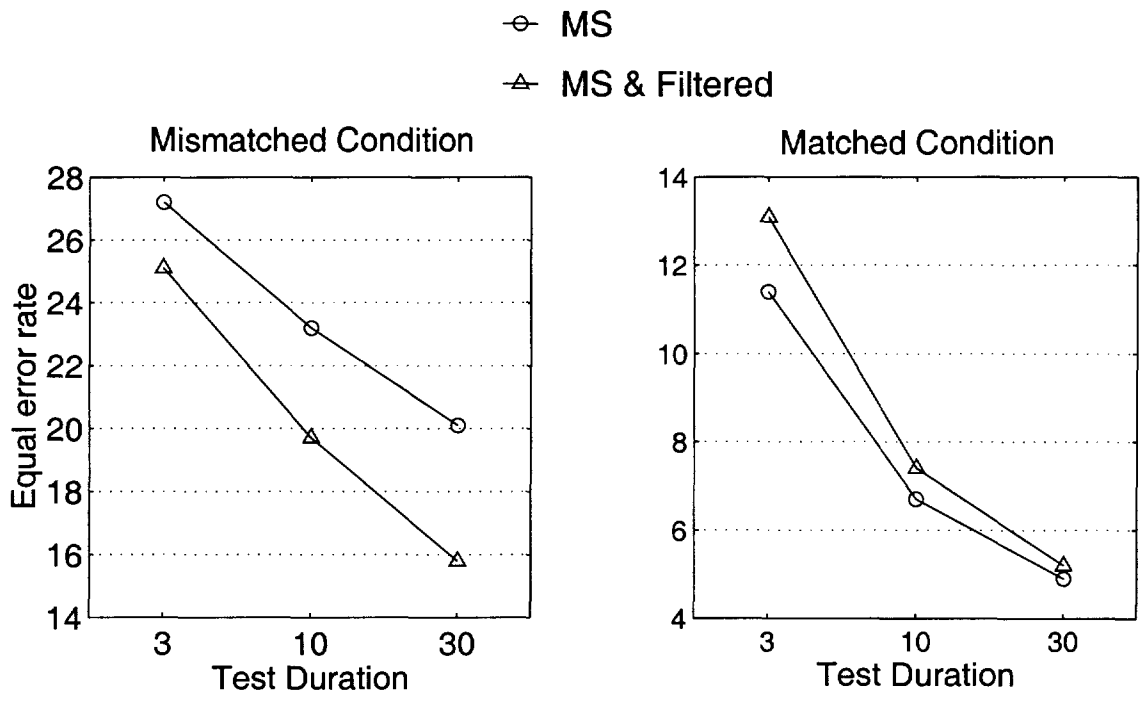


Figure 5.7: Temporal filtering can improve the performance even when applied on mean removed temporal trajectories. This illustrates the significance of removing high frequency components in the mismatched conditions for speaker verification.

high frequency cutoff. This suggests that the improvement due to the handset normalizing filter is due to the lower high-pass cutoff. Hence it can be concluded that by attenuating frequency components above 5 Hz, the mismatch between various handsets can be reduced. Note that, since the data-driven filter was designed using mean normalized feature vectors, the filter does not attenuate the DC component of the modulation spectrum and hence makes it necessary to be used in conjunction with mean subtraction.

5.6 Significance of Low Modulation Frequency Components

In this section we experimentally demonstrate the significance of low (between 0 and 1 Hz) modulation frequency components for speaker verification. For this we compare the performance of two data-driven filters: (i) one designed using mean normalized temporal

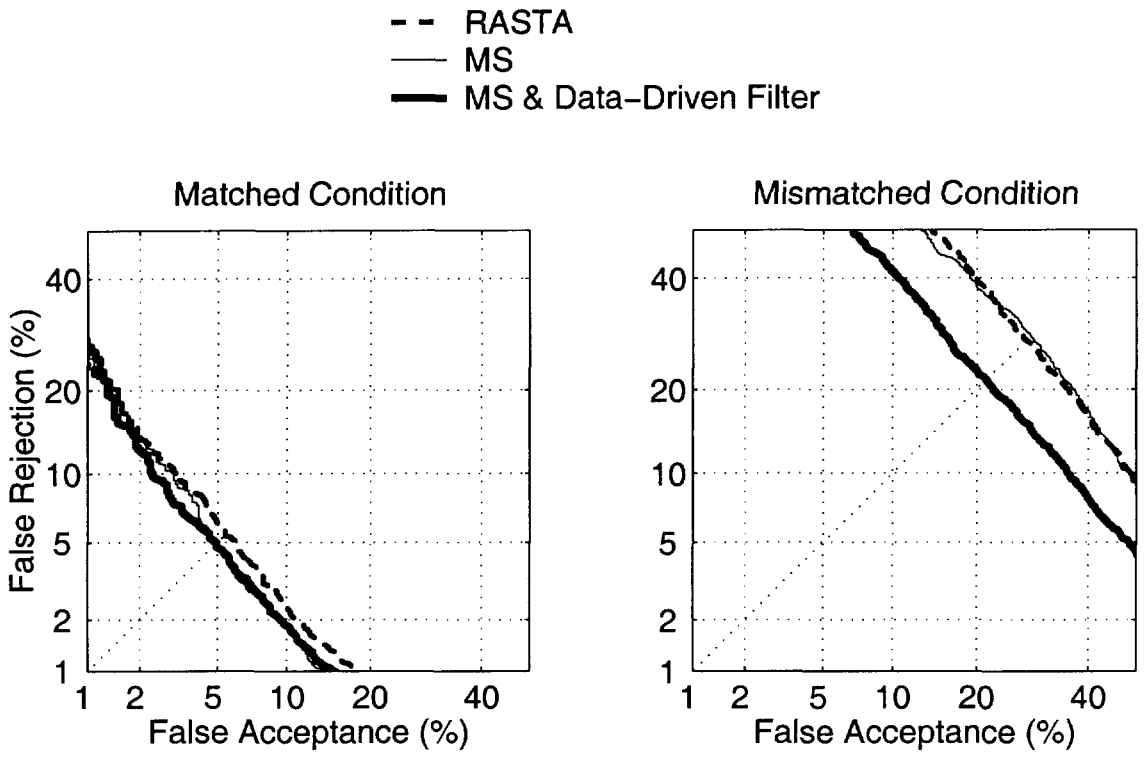


Figure 5.8: Plot illustrating the reduction in error in the mismatched condition due to data-driven temporal filtering compared to the conventional mean subtraction (MS) and RASTA filtering.

trajectories and the other (ii) designed using temporal trajectories that are not mean-normalized. Note that, since one data-driven filter is designed using mean normalized data, the filter does not attenuate the DC component of the modulation spectrum and hence this filter is used in conjunction with mean subtraction.

The performances of these two filters are compared with the RASTA filter in Figure 5.9. From the DET curve it is clear that the data-driven handset normalizing filter performs better than the RASTA filter. However its performance is worse than that obtained by the simultaneous use of the handset normalizing filter and MS. This trend can be explained as follows. The handset normalizing filter used in conjunction with MS does not attenuate frequency components below 1 Hz (broken line in Figure 5.6). The simultaneous use of this filter and MS would imply a frequency response which is dependent on the length of

Table 5.2: The performance of systems using mean subtraction, RASTA filtering and the data-driven filter. The table provides the comparison using both equal error rate (EER) and minimum decision cost function (MDCF).

Processing	Mismatched		Matched	
	EER	MDCF	EER	MDCF
MS	28.8 %	0.084	4.9 %	0.022
RASTA	27.9 %	0.086	5.4 %	0.027
MS & Data-Driven Filter	21.4 %	0.078	4.9 %	0.024

the utterance. The longer the utterance, the lesser will be the attenuation of components between 0 and 1 Hz. On the other hand, the handset normalizing filter that is used without MS attenuates the DC component (solid line in Figure 5.6). The frequency resolution of this filter is restricted to 1 Hz (since the number of taps are 101). This causes the filter to attenuate frequency components between 0 and 1 Hz irrespective of the length of the utterance. It has already been shown that modulation frequency components below 1 Hz are useful for speaker verification [75] and hence attenuation of these components causes the degradation in speaker verification performance.

5.7 Comparison with Low-pass Filtering

As discussed in the previous section, modulation frequency components below 1Hz have been shown to be important for speaker verification [1]. Earlier work by van Vuuren has shown that a low-pass filter with 10 Hz cutoff used in conjunction with mean subtraction yields improved speaker verification performance in handset mismatches [1, 75]. The proposed data-driven filter was compared with the low-pass filter in the NIST-1999 speaker recognition evaluation task.

Figure 5.10 illustrates the improvement brought about by the data-driven temporal filter over the baseline system, which uses a low pass filter. The data-driven filter causes a significant reduction in error when there is a handset type mismatch. The low-pass filter reduces the error by about 10% as compared to mean subtraction. The data-driven filter further reduces the error by another 10%. This illustrates the superiority of the

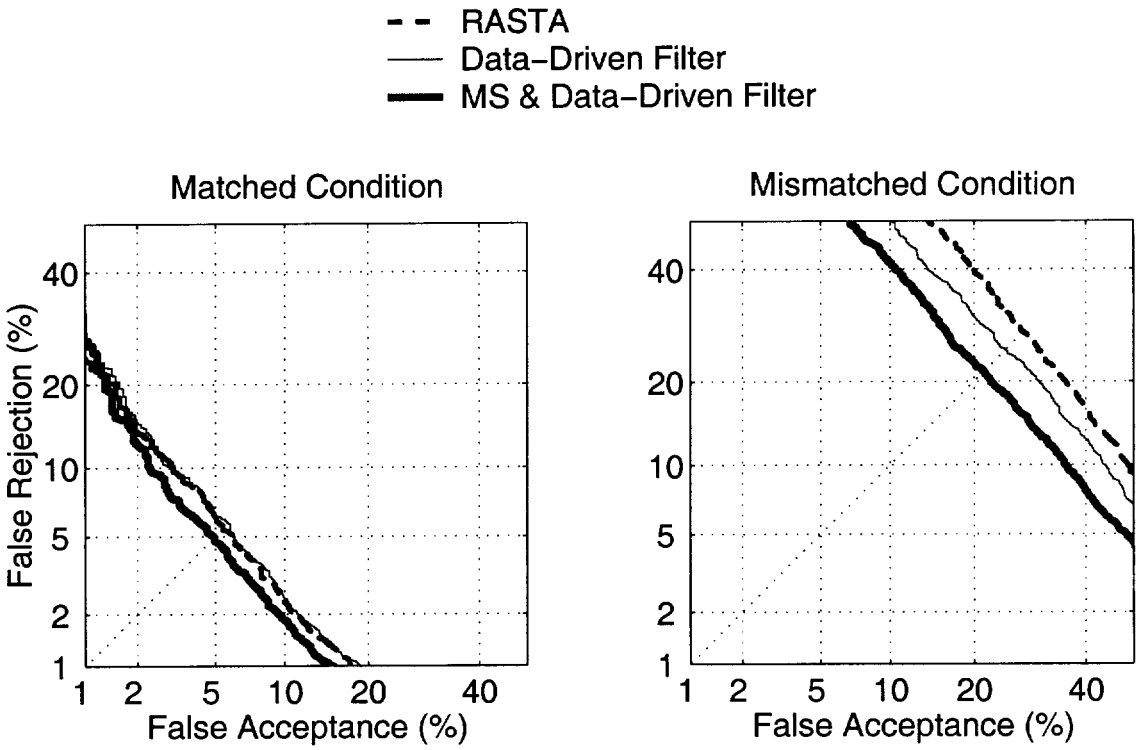


Figure 5.9: Plot illustrating the effect of mean subtraction(MS) used in conjunction with data-driven filter.

data-driven filter over the experimentally designed low-pass filter.

5.8 Summary and Comments on Temporal Filtering

In this chapter we introduced the problem of telephone handset mismatches in speaker verification. The performance of a speaker verification system is significantly degraded by the variability introduced by different telephone handsets. A handset mismatch can increase the verification error by more than five times. In speaker verification applications it is practically impossible to train the speaker models with speech from the target speaker recorded over multiple telephone handsets. The proposed data-driven filter is shown to significantly reduce speaker verification error in mismatched conditions. The proposed data-driven handset normalizing filter outperforms the conventional channel normalizing methods like RASTA filtering and mean subtraction. We have clearly demonstrated

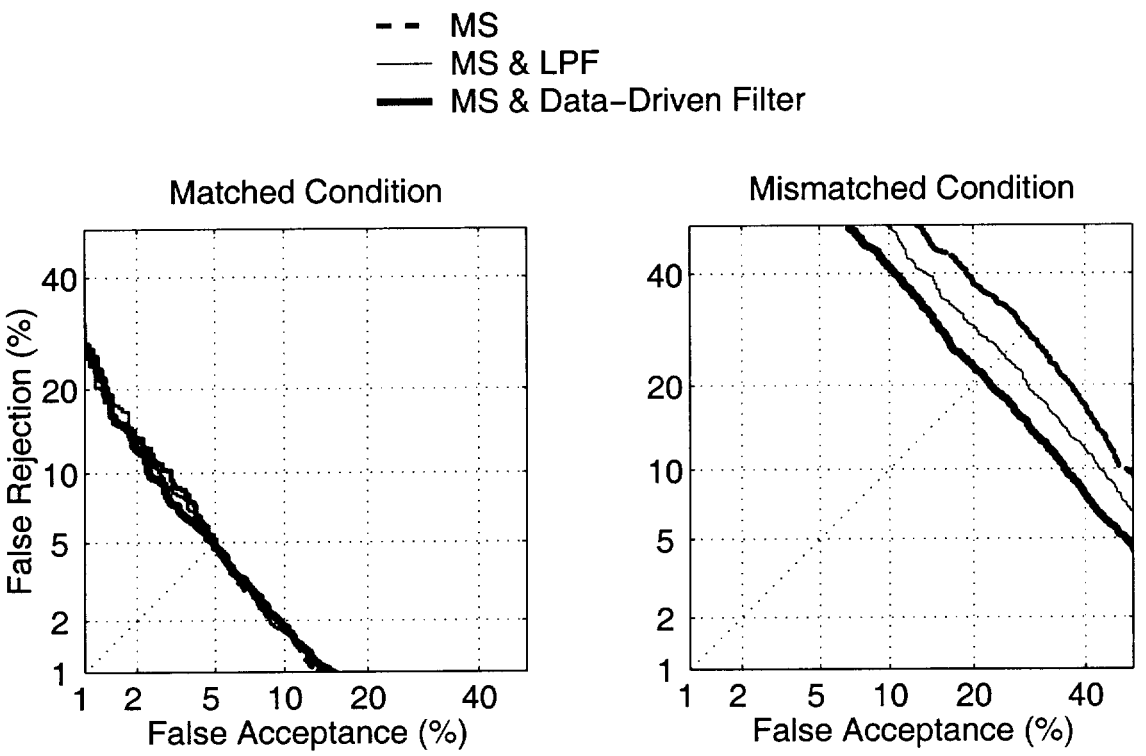


Figure 5.10: Plot comparing the performance of data-driven temporal filtering, low-pass filtering and mean subtraction.

that by using feature processing methods which are designed to alleviate unwanted variability (variability introduced by handset) the robustness of verification systems can be significantly improved. The impulse response of the handset normalizing filter can be approximated by the difference of Gaussian (DOG) function. Hence the impulse response can be represented by a small number (2) of parameters without negatively impacting performance.

Chapter 6

Summary and Future directions

The dissertation examines the hypothesis that stochastic approaches can be used to design feature extraction methods, which could provide substantial advantages over conventional feature extraction methods in the extraction of information from the speech signal. The proposed feature extractors are trained to suppress the unwanted variability while preserving the desired variability in the signal and the training is carried out using large amount of labeled speech data. Thus we call this approach data-driven. The expected advantages of this approach are the following: (a) Since feature extraction involves reduction in dimensionality, by using data-driven methods for feature extraction we are making sure that the dimensions that are preserved carry maximum amount of useful information (or carry least amount of harmful information). This could improve the performance of various recognition systems. (b) In addition to improved performance, data-driven feature extraction methods could provide knowledge about the characteristics of various sources of information carried by the speech signal. (c) It can also help in validating the optimality of the current methods for feature extraction. Data-driven methods to derive both temporal and spectral features were presented to support the above hypothesis.

The remaining part of the chapter is organized as follows. Section 6.1 summarizes the work. Section 6.2 discusses future directions and Section 6.3 concludes the dissertation with some final remarks.

6.1 Summary

In the first chapter we introduced a general framework for data-driven feature extraction. The main issues to be addressed in developing data-driven methods were discussed.

In the second chapter we examined the characteristics of the basis vectors that correspond to the direction of maximum variability derived using principal component analysis of the critical band spectrum. The results indicated a strong similarity between the PCA-derived basis functions and the cosine functions used by the DCT. Since phonemes are used as the basic units in many speech recognition systems linear discriminant analysis (LDA) of the critical-band spectrum was carried out by considering context independent phonemes as classes. The discriminant basis functions corresponding to the larger eigenvalues were found to be significantly different from the cosine functions. We also showed that the features extracted by the alternative spectral basis functions offer advantage in phoneme classification of spectral vectors. This advantage seems to hold irrespective of the complexity of the task. The LDA-derived basis functions outperformed the conventional DCT in connected digit recognition tasks as well. The zero-crossings of the LDA-derived spectral basis functions are reasonably uniformly spaced on the Bark scale of the auditory-like critical-band spectrum.

In the third chapter we investigated the nature of phonetic discriminant information present in the short-time spectrum of speech signal. The objective of such an analysis was to investigate the suitability of Mel/Bark resolution for phoneme recognition. The shape of the discriminant vectors computed using LDA determines the spectral variations that are important in discriminating phonemes. Analysis resolution of these discriminant vectors was determined by the period of a half cycle in the LDA-derived basis vectors. The smaller the period the higher the resolution. We observed that the discriminant vectors analyze the low frequency part of the spectrum with higher resolution than the high frequency part. This trend is consistent with the properties of auditory-like filters used by the conventional Mel/Bark filter-bank analysis. It was shown that this trend is prominent only while discriminating sonorant sounds. Further analysis led to the conclusion that

non-uniform width of peaks seen in the class conditional spectral means causes the discriminant functions to exhibit the property of non-uniform resolution. The sensitivity of features derived using the discriminant basis functions, to formant locations was analyzed. It was shown that the sensitivity of the features extracted by these basis functions is approximately inversely proportional to the formant frequency. This trend in sensitivity is consistent with the properties of hearing. Non-uniform frequency resolution in the form of critical-band filters is mainly used in ASR systems to emulate certain properties of human hearing. Using discriminant analysis we have established a link between non-uniform frequency resolution, phonetic discrimination and the spectral properties of vowel-like sounds.

In the fourth chapter, we proposed a method for alleviating the variability introduced by different telephone handsets. The handset normalization is achieved by filtering the time trajectories of logarithmic filter bank energies. The handset normalizing filter was designed using oriented principal component analysis (OPCA). OPCA involves the computation of noise and signal covariances in order to estimate a direction in the feature space which captures maximum signal variance while suppressing as much noise variance as possible. For designing handset normalizing filters, the variance introduced by different microphones was labeled as the noise variance. A novel method to estimate the variance introduced by handsets was also presented. OPCA is used to design a filter that optimally suppress the handset variability under the constraint of preserving phonetic variability. This method for designing filters uses only the second order statistics and thus the solution is obtained by solving a straight forward eigenvalue problem. This filter attenuates the DC component in the modulation frequency thus making the features relatively insensitive to any stationary convolutive noise. The filter emphasizes the modulation frequency components between 1 and 4 Hz. The gain of the filter drops off fairly sharply beyond 5 Hz.

In Chapter 5 we introduced the problem of telephone handset mismatches in speaker verification. A handset mismatch can increase the verification error by more than five times. In applications involving speaker verification, it is practically impossible to train the speaker models with speech from the target speaker recorded over multiple telephone

handsets. Hence the trained models are always biased to the handset used for collecting the training data. During testing, use of a different telephone handset introduces a mismatch between the statistics of the test data and the trained model, which in turn increases the probability of error. The proposed data-driven filter was shown to significantly reduce speaker verification error in mismatched conditions. The data-driven handset normalizing filter outperformed the conventional channel normalizing methods like RASTA filtering and mean subtraction. We clearly demonstrated that by using feature processing methods which are designed to alleviate undesired variability (*variability introduced by handsets*) the robustness of the verification systems can be significantly improved. It was shown that the impulse response of the handset normalizing filter can be approximated by the difference of Gaussian (DOG) function. Hence the the filter can be represented by a small number (2) of parameters without degrading the performance. We further examined the effect of using the handset normalizing filter in conjunction with mean subtraction (MS). This lead to the conclusion that preserving modulation frequency components between 0 and 1 Hz is important for speaker verification. This observation is consistent with the earlier work which investigated the importance of modulation frequency components for speaker verification.

6.2 Future Directions

The scope of the work described in this dissertation can be expanded along many directions. The following are some of the directions for future work:

- In all the feature extraction methods we have used short-time Fourier analysis to get the initial feature representation. Short-time processing uses many assumptions that are borrowed from speech coding and hence may not be optimal for speech and speaker recognition. Developing a data-driven method to extract features directly from the speech signal (without assuming short-time processing) is a challenging direction which needs to be investigated.

- The discriminant methods used to derive spectral basis functions makes a number of assumptions about the characteristics of the class-conditional distributions. If

the computational problems which are caused by relaxing this assumptions could be solved then the resulting discriminant solution could be made much more general and accurate.

- We have used a feature extraction model given by, $\mathbf{f} = \mathbf{F}(\mathbf{s})$, where $\mathbf{F}(\cdot)$ is the feature extraction function, $\mathbf{s}$ is the initial representation and $\mathbf{f}$ is the extracted feature. In the above feature extraction model, we have assumed that the function $\mathbf{F}(\cdot)$ as linear. This assumption makes the feature extractor sub-optimal. A generalized non-linear model, if trained with enough amount of data could be much more powerful than the linear models discussed in this dissertation.

- For designing spectral basis functions we have assumed phonemes as the classes. While phonemes are the smallest segments of sound that can cause a change in the meaning of a word, it may not be the optimal sub-word unit for automatic speech recognition. Methods for automatically deriving the optimal set of sub-word units from data is another important aspect which we have not explored in this work.

5.1 Final Comments

The work presented in this dissertation represents an effort to optimize feature extraction methods for speech recognition and speaker verification. The majority of the current feature extraction modules use assumptions which are not validated as optimal for pattern recognition. We have introduced a general paradigm for data-driven feature extraction that attempts to derive knowledge from speech data and this knowledge is used to design better feature extraction methods. We have presented two specific data-driven feature extraction methods, (i) spectral basis functions for speech recognition and (ii) temporal filters for speaker verification. These feature extraction methods demonstrated the advantages of data-driven feature extraction. The data-driven methods presented in this dissertation, in spite of providing significant improvement in the performance of speaker verification and speech recognition, are sub-optimal due to the assumptions that were made while designing. Future work should be directed towards making these assumptions

realistic which would lead to more effective realization of data-driven feature extraction modules. Finally, we also hope that the data-driven approach for analyzing the various sources of information in the speech signal will lead us to a better understanding of speech production, transmission and perception.

Appendix A

List of phonemes

Table A.1: The phonetic categories considered for many of the experiments reported in the dissertation.

Broad class	Phoneme	Example	Broad class	Phoneme	Example
Vowels	iy	beet	Stops	k	can
	eh	bet		b	ban
	ih	bit		g	gander
	ae	bat		d	dan
	ux	suit		t	tan
	ah	above	Fricatives	f	fine
	ao	caught		s	sign
	aa	father		v	vine
	ix	roses		z	resign
	ax	above		th	thigh
	uh	book	Nasals	m	me
	uw	boot		n	knee
Diphthongs and Glides	ay	bye			
	ey	bay			
	ow	boat			
	l	lent			
	r	rent			
	y	yes			
	w	went			

Appendix B

Mean Spectral Patterns

In this Appendix the shape of the mean spectral vectors corresponding to various phonemes are illustrated. The spectra were computed using a 256 point DFT. The spectral means corresponding to 36 phonemes were computed from 538471 frames of data. Table B.1 shows the number of frames used for the computation of each of the 36 phonetic means. Figure B.1 and Figure B.2 shows the mean spectral patterns of these sounds.

Table B.1: The number of frames used for the computation of spectral means of various phonemes.

Phoneme	Examples	Phonemes	Examples
iy	30821	eh	19375
ih	32108	ae	23934
ux	0	ah	43372
ao	5248	aa	20118
ix	3518	ax	7980
uh	2137	uw	9938
ay	25952	ey	18641
ow	15311	l	19771
r	17390	y	3368
w	12626	f	15403
v	8028	th	4541
s	46175	z	12696
kcl	14290	bcl	7556
gcl	3126	dcl	10923
tcl	22299	k	10581
b	1290	g	1474
d	3379	t	15897
m	18067	n	31138

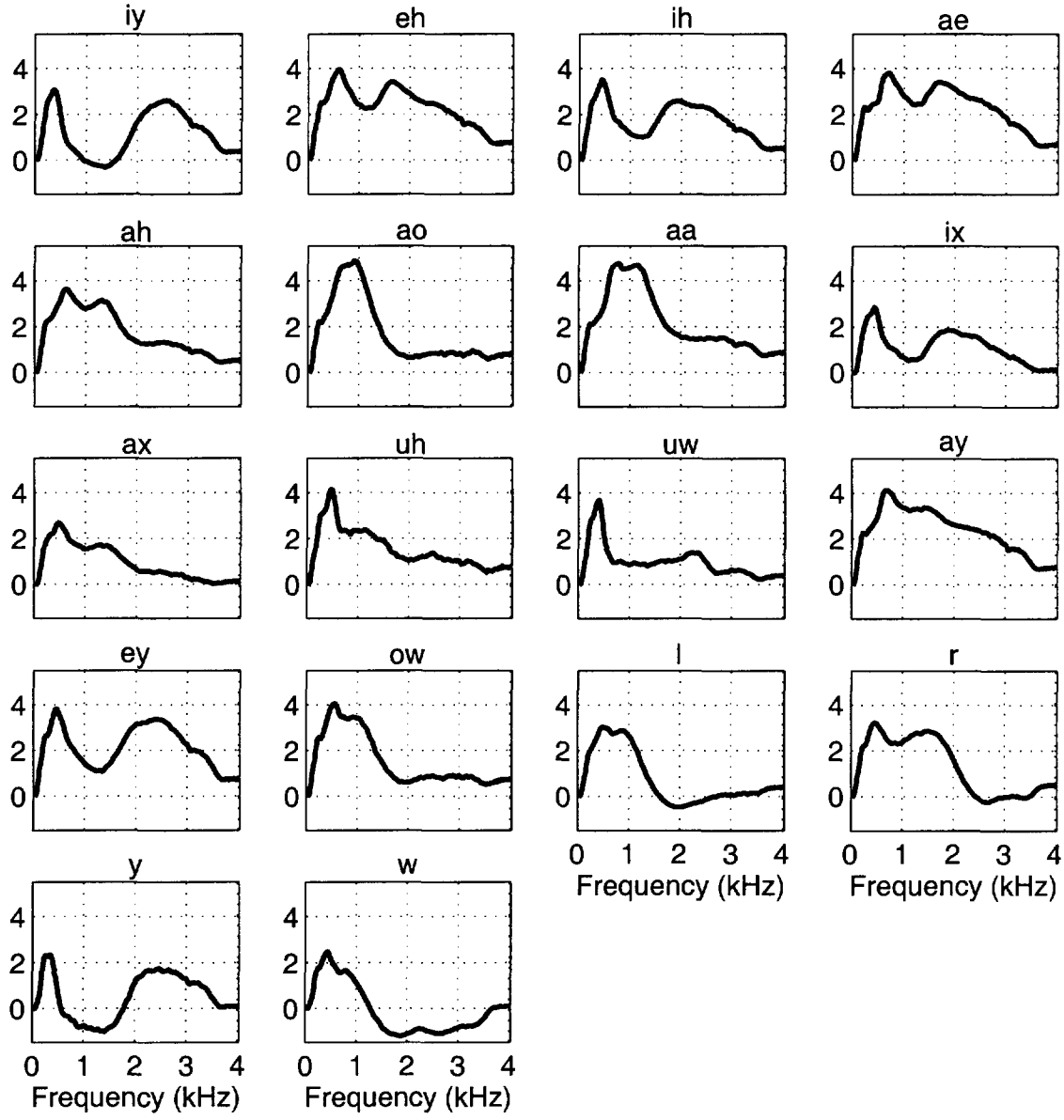


Figure B.1: The mean spectral patterns of vowels, diphthongs and glides computed from the OGI STORIES corpus.

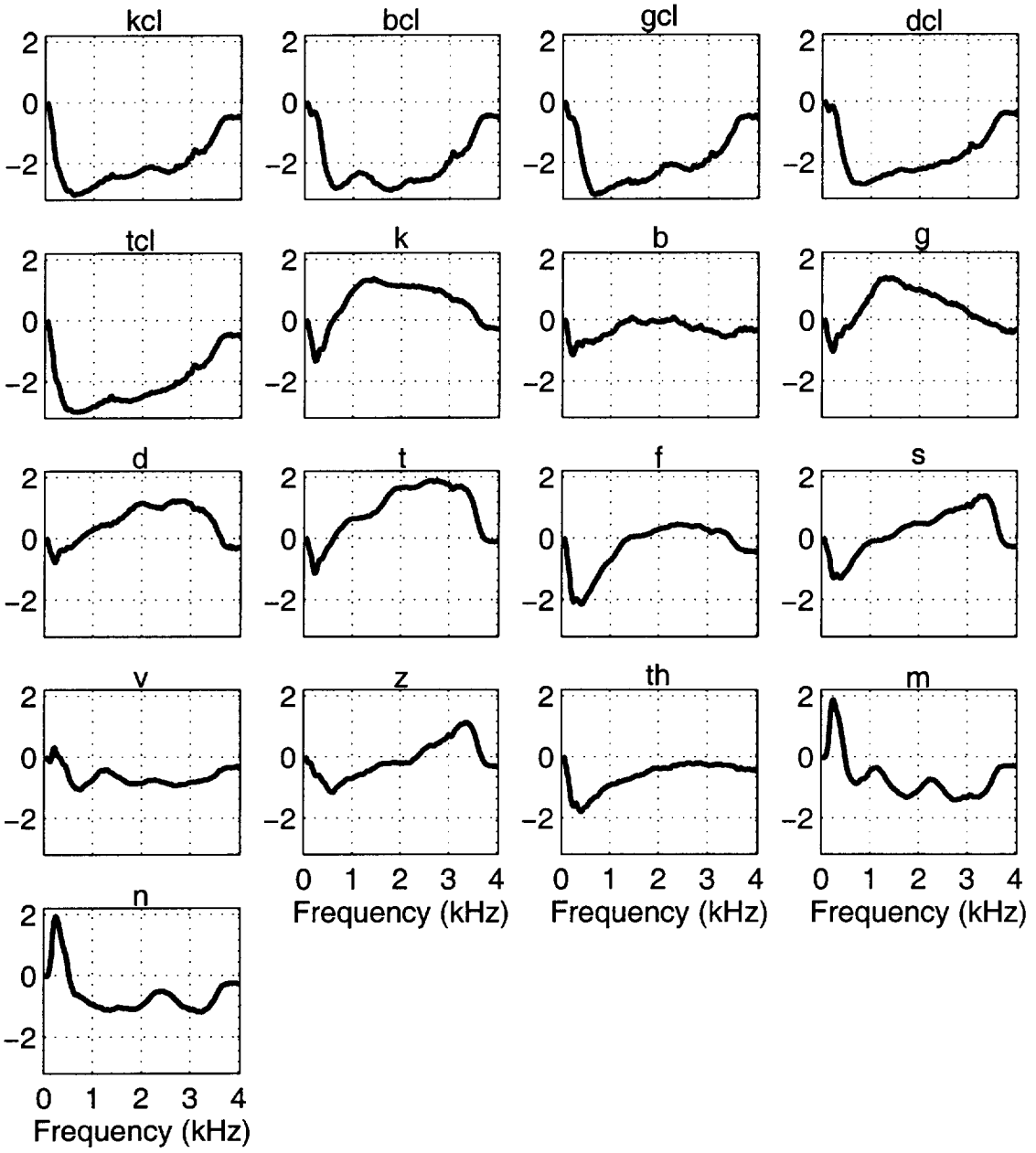


Figure B.2: The mean spectral patterns of stops, fricatives and nasals computed from the OGI STORIES corpus.

Appendix C

Difference of Means as Discriminants

In this appendix we derive the conditions under which the difference vector between two class conditional means becomes the discriminant vector. If the two classes have equal prior probabilities then the decision rule which yields minimum error is given by the likelihood ratio test

$$\ln \frac{p(\lambda_1|x)}{p(\lambda_2|x)} \underset{\lambda_2}{\overset{\lambda_1}{\lessgtr}} 0, \tag{C.1}$$

where the two classes in which we are interested are λ_1 and λ_2 and x is the vector of measurements. The class conditional distributions are assumed to be normal with means given by μ_1 and μ_2 respectively. The class conditional covariances are given by $\Sigma_1 = \Sigma_2 = I$. Under these assumptions the log-likelihood ratio is given by

$$\begin{aligned} \ln \frac{p(\lambda_1|x)}{p(\lambda_2|x)} &= (x-\mu_2)^T(x-\mu_2) - (x-\mu_1)^T(x-\mu_1), \\ &= \mu_1^T\mu_1 - \mu_2^T\mu_2 + 2x^T(\mu_1-\mu_2). \end{aligned} \tag{C.2}$$

Let $d = \mu_1 - \mu_2$, then, from equation C.1 and C.2 the modified log-likelihood ratio test can be written as

$$2x^T d \underset{\lambda_2}{\overset{\lambda_1}{\lessgtr}} \frac{1}{2}(|\mu_2|^2 - |\mu_1|^2) \tag{C.3}$$

The feature extraction using the discriminant direction d, reduces to projecting the unknown measurement vector x, onto d which yields a feature $f = x^T d$.

From the above analysis it is clear that the feature f computed by projecting the measurement x onto d can be directly compared to a threshold to perform the classification. This illustrates the significance of the discriminant vector defined by the difference between the two class conditional means.

Appendix D

Conditions for Non-uniform resolution

In this appendix we investigate the conditions under which discriminant basis vectors exhibits non-uniform analysis resolution. This analysis is carried out by approximating the mean spectral patterns of sonorant speech sounds by Gaussian functions.

Sonorants are mainly characterized by the vocal tract resonances, which in turn are manifested in the speech signal as spectral peaks. Hence it should be interesting to see how the simple discriminant vector d discriminates two spectral peaks which are displaced in frequency. The two spectral peaks are assumed to have the same width and are approximated by the Gaussian form, $N_{p,b}(f) = e^{\frac{(f-p)^2}{b}}$. $N_{p,b}(f)$ is a function of f the frequency and is parameterized by p, the location of the spectral peak and the parameter b, which controls the width of the spectral peak. The two class conditional means μ_1 and μ_2 are represented by $N_{p1,b}(f)$ and $N_{p2,b}(f)$ respectively. The discriminant function d is thus given by the difference of two Gaussian functions (DOG) with different means and same variance. Figure D.1 illustrates the shape of the discriminant function d. The analysis resolution of the DOG discriminant function is dependent on the parameter b which corresponds to a bandwidth-like measure of the two class-conditional mean vectors. The increased analysis resolution is caused by a smaller value of b which in turn implies a narrower bandwidth. It is clear from the figure that the analysis resolution is inversely proportional to the bandwidth of the two spectral peaks to be discriminated. Hence it can be concluded that if the bandwidth of the spectral peaks that appear in the class conditional means increases with the location of the center frequency, then the discriminant function will have non-uniform analysis resolution. In particular the discriminant function will show high analysis resolution in low frequency regions and low resolution in

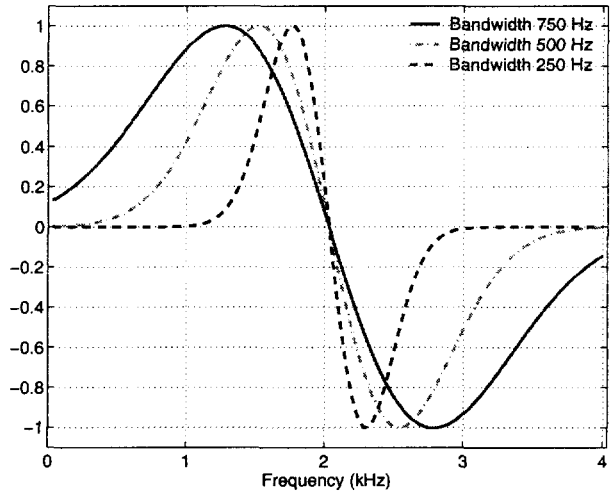


Figure D.1: Difference of Gaussians(DOG) as a function of the variance (width of the spectral peak). If spectral peaks are modeled by Gaussians then DOG becomes the discriminant functions to classify spectral peaks which are shifted in frequency. The figure indicates that the analysis resolution of these discriminant functions are inversely proportional to the width associated with the spectral peak.

high frequency region.

The two class problem is now made more realistic by modeling the class conditional means as a linear combination of two Gaussian functions described by

$$\mu_1 = N_{p_1,b_1}(f) + N_{p_2,b_2}(f), \tag{D.1}$$
$$\mu_2 = N_{p_1+\delta_1,b_1}(f) + N_{p_2+\delta_2,b_2}(f).$$

The class conditional mean μ_1 consists of two spectral peaks located at p_1 and p_2 with bandwidths of b_1 and b_2 respectively. The mean vector corresponding to the competing class has the same two spectral peaks displaced by δ_1 and δ_2 in frequency. These characteristic s are typical of two distinct vowel-like sounds. The discriminant function corresponding to these two classes is then given by the difference of the two class conditional means. This discriminant vector will thus exhibit progressively decreasing analysis resolution if $b_1 < b_2$. Figure D.2 shows the effect of bandwidth increase in making the analysis resolution of the discriminant vectors non-uniform.

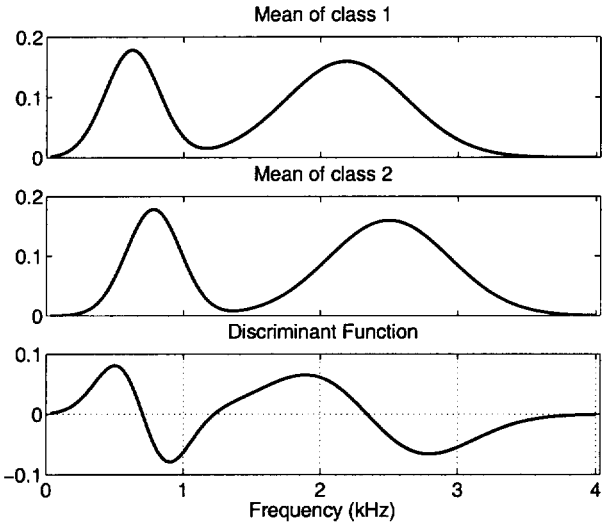


Figure D.2: The shape of the discriminant function clearly indicates a higher analysis resolution in the lower frequency region and a lower analysis resolution in the high frequency region. This is an effect of higher bandwidth associated the spectral peak which is located at higher frequency.

Bibliography

[1] Sarel van Vuuren and Hynek Hermansky, "On the importance of components of the modulation spectrum for speaker verification," in *Proceedings of International Conference on Spoken Language Processing*, Sydney, Australia, 1998, pp. 3205–3208.

[2] L. Rabiner and B. Juang, *Fundamentals of Speech Recognition*, Prentice Hall, First edition, 1993.

[3] Bishnu S. Atal, "Automatic recognition of speakers from their voices," *Proceedings of the IEEE*, vol. 64, pp. 460–475, 1974.

[4] D. O'Shaughnessy, "Speaker recognition," *IEEE Acoustics, Speech and Signal Processing Magazine*, vol. 3, pp. 4–17, 1986.

[5] G. R. Doddington, "Speaker recognition - Identifying people by their voices," *Proceedings of the IEEE*, vol. 11, pp. 1651–1664, 1985.

[6] H. Gish and M. Schmidt, "Text-independent speaker identification," *IEEE Signal Processing Magazine*, vol. 11, no. 4, pp. 18–32, 1994.

[7] Douglas A. Reynolds, Thomas F. Quatieri, and Robert B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, In press for 2000.

[8] Chin-Hui Lee, Chih-Heng Lin, and Biing-Hwang Juang, "A study on speaker adaptation of the parameters of continuous density hidden Markov models," *IEEE Transactions on Signal Processing*, vol. 39, no. 4, pp. 806–814, 1991.

[9] Jun-ichi Takahashi and Shigeki Sagayama, "Vector-field-smoothed Bayesian learning for fast and incremental speaker/telephone-channel adaptation," *Computer Speech & Language*, vol. 11, pp. 127–146, 1997.

[10] J. W. Picone, "Signal modeling techniques in speech recognition," *Proceedings of the IEEE*, vol. 81, no. 9, pp. 1215–1247, 1993.

[11] B. S. Atal and S. L. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave," *The Journal of the Acoustical Society of America*, vol. 50, pp. 637–655, 1971.

[12] P. Mermelstein, "Distance measures for speech recognition, psychological and instrumental," in *Pattern Recognition and Artificial Intelligence*, R. C. H. Chen, Ed., pp. 374–388. Academic Press, 1976.

[13] Steven B. Davis and Paul Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," in *Pattern Recognition and Artificial Intelligence*, Alex Waibel and Kai-Fu Lee, Eds., pp. 65–74. Morgan Kaufmann Publishers, 1990.

[14] Hynek Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *The Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, 1990.

[15] Peter Ladefoged, *A Course in Phonetics*, Harcourt Brace Jovanovich, Second edition, 1982.

[16] M. Narendranath, Hema A. Murthy, S. Rajendran, and B. Yegnanarayana, "Transformation of formants for voice conversion using artificial neural networks," *Speech Communication*, vol. 16, pp. 207–216, 1995.

[17] Athanasious Papoulis, *Probability, Random Variables and Stochastic Processes*, McGraw-Hill, Third edition, 1991.

[18] Thomas M. Cover and Joy A. Thomas, *Elements of Information Theory*, John Wiley & Sons, First edition, 1991.

[19] F. Jelinek, *Statistical Methods for Speech Recognition*, MIT Press, First edition, 1997.

[20] Douglas A. Reynolds, "Speaker identification and verification using Gaussian mixture models," *Speech Communication*, vol. 17, pp. 91–108, 1995.

[21] Douglas A. Reynolds, "Comparison of background normalization methods for text-independent speaker verification," in *Proceedings of the European Conference on Speech Communication and Technology*, Rhodes, Greece, 1997, pp. 963–966.

[22] Alvin Martin et al., Eds., *Speaker Recognition Workshop Notebook*, NIST, University of Maryland, College Park, MD, March/April, 1998.

[23] Hynek Hermansky and Nelson Morgan, "RASTA processing of speech," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4, pp. 578–589, 1995.

[24] Howard Yang, Sarel van Vuuren, and Hynek Hermansky, "Relevancy of time-frequency features for phonetic classification measured by mutual information," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, Phoenix, USA, 1999, pp. 225–228.

[25] J. A. Bilmes, "Maximum mutual information based reduction strategies for cross-correlation based joint distribution modeling," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, Seattle, USA, 1998, pp. 469–472.

[26] Narendranath Malayath, Hynek Hermansky, Sachin Kajarekar, and B. Yegnanarayana, "Data-driven temporal filters and alternatives to GMM in speaker verification," *Digital Signal Processing*, In press for 2000.

[27] Carlos Avendano, *Temporal Processing of Speech in a Time-Feature Space*, Ph.D., Thesis, Department of Electrical and Computer Engineering, Oregon Graduate Institute of Science and Technology, 1997.

[28] Carlos Avendano, Sarel Van Vuuren, and Hynek Hermansky, "Data based filter design for RASTA-like channel normalization in ASR," in *Proceedings of the International Conference on Spoken Language Processing*, Philadelphia, USA, 1996, pp. 2087–2090.

[29] Sarel van Vuuren and Hynek Hermansky, "Data-driven design of RASTA-like filters," in *Proceedings of European Conference on Speech Communication and Technology*, Rhodes, Greece, 1997, pp. 409–412.

[30] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of Eugenics*, vol. 7, pp. 179–188, 1936.

[31] R. A. Fisher, "The statistical utilization of multiple measurements," *Annals of Eugenics*, vol. 8, pp. 376–386, 1938.

[32] Hynek Hermansky and Narendranath Malayath, "Spectral basis functions from discriminant analysis," in *Proceedings of International Conference on Spoken Language Processing*, Sydney, Australia, 1998, pp. 1379–1382.

[33] V. R. Lesser, R. D. Fennel, L. D. Erman, and D. R. Reddy, "Organization of the Hearsay-II speech understanding system," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 23, no. 1, pp. 11–23, 1975.

[34] D. Raj Reddy, "Speech recognition by machine: A review," *Proceedings of the IEEE*, vol. 64, pp. 502–531, 1976.

[35] Victor W. Zue, "The use of speech knowledge in automatic speech recognition," in *Pattern Recognition and Artificial Intelligence*, Alex Waibel and Kai-Fu Lee, Eds., pp. 200–213. Morgan Kaufmann Publishers, 1990.

[36] G. Mercier, D. Bigorgne, L. Miclet, and M. Querre, "Recognition of speaker-dependent continuous speech with KEAL," in *Pattern Recognition and Artificial Intelligence*, Alex Waibel and Kai-Fu Lee, Eds., pp. 225–234. Morgan Kaufmann Publishers, 1990.

[37] Dennis H. Klatt, "Review of the ARPA speech understanding project," *The Journal of the Acoustical Society of America*, vol. 62, pp. 1345–1366, 1977.

[38] Nelson Morgan and Herve A. Bourlard, "Neural networks for statistical recognition of continuous speech," *Proceedings of the IEEE*, vol. 83, pp. 742–770, 1995.

[39] S. Katz, "Estimation of probabilities from sparse data for language model component of a speech recognizer," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 35, pp. 400–401, 1987.

[40] Hynek Hermansky, "Should recognizers have ears ?," *Speech Communication*, vol. 25, pp. 3–27, 1998.

[41] Melvyn Hunt, "A statistical approach to metrics for word and syllable recognition," *The Journal of the Acoustical Society of America*, vol. 66(S1), pp. 35–36, 1979.

[42] Ronald W. Schafer and Lawrence R. Rabiner, "Digital representations of speech signals," in *Pattern Recognition and Artificial Intelligence*, Alex Waibel and Kai-Fu Lee, Eds., pp. 49–64. Morgan Kaufmann Publishers, 1990.

[43] Alan V. Oppenheim and Ronald W. Schafer, *Digital Signal Processing*, Prentice-Hall Inc., First edition, 1975.

[44] E. Zwicker, G. Flottorp, and S. S. Stevens, "Critical bandwidth in loudness summation," *The Journal of the Acoustical Society of America*, vol. 29, pp. 548–557, 1957.

[45] A. M. Noll, "Cepstrum pitch determination," *The Journal of the Acoustical Society of America*, vol. 41, pp. 293–309, 1967.

[46] Alan V. Oppenheim and R. W. Schafer, "Homomorphic analysis of speech," *IEEE Transactions on Audio and Electroacoustics*, vol. 16, pp. 221–226, 1968.

[47] A. K. Jain, "A sinusoidal family of unitary transforms," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 1, pp. 356–365, 1979.

[48] N. Merhav and Chin-Hui Lee, "On the asymptotic statistical behavior of empirical cepstral coefficients," *IEEE Transactions on Signal Processing*, vol. 41, no. 3, 1993.

[49] K. Fukunaga, *Statistical Pattern Recognition*, Academic Press, Second edition, 1990.

[50] Melvyn Hunt and Claude Lefebvre, "Speaker dependent and independent speech recognition experiments with an auditory model," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, New York, USA, 1988, pp. 215–218.

[51] Melvyn Hunt and Claude Lefebvre, "A comparison of several acoustic representations for speech recognition with degraded and undegraded speech," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, Glasgow, UK, 1989, pp. 262–265.

[52] Melvyn Hunt, Stephen Richardson, Charies Bateman, and Alain Piau, "An investigation of PLP and IMELDA acoustic representations and of their potential for combination," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, Toronto, Canada, 1991, pp. 881–884.

[53] C. M. Ayer, M. J. Hunt, , and D. M. Brookes, "A discriminatively derived linear transformation for improved speech recognition," in *Proceedings of European Conference on Speech Communication and Technology*, Berlin, Germany, 1993, pp. 583–586.

[54] P. Brown, *The Acoustic-Modeling Problem in Automatic Speech Recognition*, Ph.D., Thesis, Carnegie Mellon University, 1987.

[55] G. R. Doddington, "Phonetically sensitive discriminants for improved speech recognition," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, Glasgow, UK, 1989, pp. 556–559.

[56] P. C. Woodland and D. R. Cole, "Optimizing hidden Markov models using discriminative output distributions," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, Toronto, Canada, 1991, pp. 545–548.

[57] E. Gunter Schurt-Talamazzini, Joachim Hornegger, and Heinrich Niemann, "Optimal linear transformations for semi-continuous Hidden Markov Models," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, Detroit, USA, 1995, pp. 369–372.

[58] T. Eisele, R. Haeb-Umbach, and D. Langmann, "A comparative study of linear feature transformation techniques for automatic speech recognition," in *Proceedings of International Conference on Spoken Language Processing*, Philadelphia, USA, 1996, pp. 252–255.

[59] E. Batlle, C. Nadeu, and J. A. R. Fonollosa, "Feature decorrelation methods in speech recognition. A comparative study," in *Proceedings of International Conference on Spoken Language Processing*, Sydney, Australia, 1998, pp. 951–954.

[60] Don X. Sun and Li Deng, "Analysis of acoustic-phonetic variations in fluent speech using TIMIT," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, Detroit, USA, 1995, pp. 201–204.

[61] Narendranath Malayath, Hynek Hermansky, and Alex Kain, "Towards decomposing the sources of variability in speech," in *Proceedings of the European Conference on Speech Communication and Technology*, Greece, 1997, pp. 497–500.

[62] Sachin Kajarekar, Narendranath Malayath, and Hynek Hermansky, "Analysis of sources of variability in speech," in *Proceedings of European Conference on Speech Communication and Technology*, Budapest, Hungary, 1999, pp. 343–346.

[63] D. H. Kil and F. B. Shin, *Pattern Recognition and Prediction with Applications to Signal Characterization*, American Institute of Physics, 1996.

[64] Shihab S. Shamma and Kung Sang Wang, "Spectral shape analysis in the central auditory system," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 5, pp. 382–394, 1995.

[65] Sadaoki Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 29, pp. 254–272, 1981.

[66] R. L. Wegel and C. E. Lane, "The auditory masking of one sound by another and its probable relation to the dynamics of inner ear," *Physical Review*, vol. 23, pp. 266–285, 1924.

[67] H. Fletcher, "Auditory patterns," *Review of Modern Physics*, vol. 12, pp. 47–65, 1940.

[68] P. M. Hamilton, "Noise masked thresholds as a function of tonal duration and masking noise band width," *The Journal of the Acoustical Society of America*, vol. 29, pp. 506—511, 1957.

[69] D. D. Greenwood, "Auditory masking and the critical band," *The Journal of the Acoustical Society of America*, vol. 33, pp. 484–501, 1961.

[70] M. R. Schroeder, "Recognition of complex acoustic signals," in *Life Sciences Research Report - 5*, T. H. Bullock, Ed., p. 324. Abkon Verlag, 1977.

[71] S. S. Stevens and J. Volkmann, "The relation of pitch of frequency: A revised scale," *The American Journal of Psychology*, vol. 53, pp. 329–353, 1940.

[72] S. Umesh, L. Cohen, and D. Nelson, "Frequency-warping and speaker-normalization," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, Munich, Germany, 1997, pp. 983–987.

[73] J. Flanagan, "Difference limen for vowel formant frequency," *The Journal of the Acoustical Society of America*, vol. 27, pp. 613–617, 1955.

[74] Alvin Martin et al., Eds., *Speaker Recognition Workshop Notebook*, NIST, Maritime Institute of Technology, Lithicum Heights MD, June, 1997.

[75] Sarel van Vuuren, *Speaker recognition in a Time-Feature Space*, Ph.D., Thesis, Department of Electrical and Computer Engineering, Oregon Graduate Institute of Science and Technology, 1999.

[76] T. Stockham, T. Cannon, and R. Ingebretsen, "Blind deconvolution through digital signal processing," *Proceedings of the IEEE*, vol. 63, pp. 678–692, 1975.

[77] R. Schwarz, T. Anastasakos, F. Kubala, J. Makhoul, L. Nguyen, and G. Zavaliagkos, "Comparative experiments on large vocabulary speech recognition," in *Proceedings of ARPA Workshop on Human Language Technology*, Plainsboro, USA, 1993.

[78] Alvin Martin et al., Eds., *Speaker Recognition Workshop Notebook*, NIST, University of Maryland, College Park, MD, May, 1999.

[79] Mark A. Przybocki and Alvin F. Martin, "NIST speaker recognition evaluation -1997," in *Proceedings of Speaker Recognition and its Commercial and Forensic Applications*, Avignon, France, 1998, pp. 120–123.

[80] T. F. Quatieri, D. A. Reynolds, and G. C. O'Leary, "Magnitude-only estimation of handset nonlinearity with application to speaker recognition," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, Seattle, USA, 1998, pp. 745–748.

[81] Douglas A. Reynolds, "HTIMIT and LLHDB: Speech corpora for the study of handset transducer effects," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, Munich, 1997, pp. 1535–1538.

[82] L. F. Lamel, R. H. Kasel, and S. Seneff, "Speech database development: Design and analysis of the acoustic-phonetic corpus," in *Proceedings of the DARPA Speech Recognition Workshop*, 1987, pp. 26–32.

[83] K. I. Diamantaras and S. Y. Kung, *Principal Component Neural Networks - Theory and Applications*, John Wiley & Sons, First edition, 1996.

[84] Hynek Hermansky and Sangita Sarma, "TRAPS - classifiers of temporal patterns," in *Proceedings of International Conference on Spoken Language Processing*, Sydney, Australia, 1998, pp. 1003–1006.

[85] Sangita Sarma, *Multi-Stream approach to Robust Speech Recognition*, Ph.D., Thesis, Department of Electrical and Computer Engineering, Oregon Graduate Institute of Science and Technology, 1999.

[86] Sarel van Vuuren and Hynek Hermansky, "Data-driven design of RASTA-like filters," in *Proceedings of the European Conference on Speech Communication and Technology*, Rhodes, Greece, 1997, pp. 409–412.

[87] Alvin Martin and Mark Przybocki, "The NIST 1999 speaker recognition evaluation - an overview," *Digital Signal Processing*, In press for 2000.

[88] Sadaoki Furui, "Speaker-independent isolated word recognition based on emphasized spectral dynamics," in *Proceedings International Conference on Acoustics, Speech and Signal Processing*, Tokyo, Japan, 1986, pp. 1991–1994.

[89] Sarel van Vuuren and Hynek Hermansky, "!MESS, a modular efficient speaker verification system," in *Proceedings of Speaker Recognition and its Commercial and Forensic Applications*, Avignon, France, 1998, pp. 198–201.

[90] J. Ortega and W. Rheinbolt, *Iterative Solutions of non-linear equations in several variables*, Academic Press, First edition, 1970.

[91] H. Ney, "Statistical modeling and dynamic programming in speech recognition," *Sprache unt Datenverarbeitung*, vol. 8, pp. 17–33, 1984.

Biographical Note

Narendranath Malayath was born in Palakkad, Kerala, India, on the 15th of May, 1971. He completed the Bachelor of Technology degree in 1992 from the NSS college of engineering, Palakkad. He then went on to work as a project associate at the Speech and Vision Laboratory, Indian Institute of Technology (IIT), Madras, India. It was from here he was inspired by Prof. B. Yegnanarayana and joined the speech research program at the computer science department. He received the M S degree from the Indian Institute of Technology in 1995. During his stay at IIT he designed and implemented a voice conversion system.

From 1995 to 1996 he worked as a software engineer at the Tata Information Systems (now IBM Global Services) where he worked on text-dependent speaker verification.

In early 1997 he joined the Oregon Graduate Institute as a Ph.D. student and became a member of the Anthropic Signal Processing Group. At OGI, under the guidance of Prof. Hynek Hermansky, he worked on several research projects which include, development of a speaker verification system based on speaker-specific mapping and design of robust feature extraction schemes. During the summer of 1998 he worked as a summer intern at the Media Technologies Laboratory at Texas Instruments under the guidance of Dr. Lorin Netsch where he developed adaptation methods for command phrase recognition in cellular phones. He has authored several conference and journal papers.

Selected Publications

1. Narendranath Malayath, Hynek Hermansky, Sachin Kajarekar, and B. Yegnanarayana, "Data-driven temporal filters and alternatives to GMM in speaker verification", *Digital Signal Processing*, 2000, In press.

2. Hynek Hermansky and Narendranath Malayath, "Spectral basis functions from discriminant analysis", In *Proceedings of International Conference on Spoken Language Processing*, pages 1379–1382, Sydney, Australia, 1998.

3. Sachin Kajarekar, Narendranath Malayath, and Hynek Hermansky, "Analysis of sources of variability in speech", In *Proceedings of European Conference on Speech Communication and Technology*, pages 343–346, Budapest, Hungary, 1999.

4. Narendranath Malayath, Hynek Hermansky, and Alex Kain, "Towards decomposing the sources of variability in speech", In *Proceedings of the European Conference on Speech Communication and Technology*, pages 497–500, Greece, 1997.

5. M. Narendranath, Hema A. Murthy, S. Rajendran, and B. Yegnanarayana, "Transformation of formants for voice conversion using artificial neural networks", *Speech Communication*, 16:207–216, 1995.