RL-TR-95-257
Final Technical Report
December 1995

COMMUNICATIONS CHANNEL NORMALIZATION TECHNIQUES

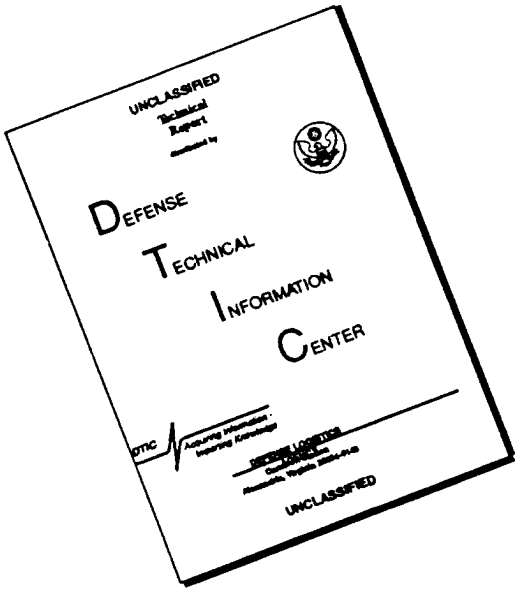
Rutgers University

Dr. Devang Naik and Dr. Richard Mammone

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

Rome Laboratory
Air Force Materiel Command
Rome, New York

DISCLAIMER NOTICE



THIS DOCUMENT IS BEST QUALITY AVAILABLE. THE COPY FURNISHED TO DTIC CONTAINED A SIGNIFICANT NUMBER OF PAGES WHICH DO NOT REPRODUCE LEGIBLY.

RL-TR-95-257 has been reviewed and is approved for publication.

APPROVED: John Grieco

JOHN GRIECO
Project Engineer

FOR THE COMMANDER:

JOSEPH CAMERA
Technical Director
Intelligence & Reconnaissance Directorate

DESTRUCTION NOTICE - For classified documents, follow the procedures in DOD 5200.22M, Industrial Security Manual or DOD 5200.1-R, Information Security Program Regulation. For unclassified limited documents, destroy by any method that will prevent disclosure of contents or reconstruction of the document.

If your address has changed or if you wish to be removed from the Rome Laboratory mailing list, or if the addressee is no longer employed by your organization, please notify Rome Laboratory/ (IRAA), Rome NY 13441. This will assist us in maintaining a current mailing list.

Do not return copies of this report unless contractual obligations or notices on a specific document require that it be returned.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY (Leave Blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	December 1995	Final Jan 94 - Jan 95

4. TITLE AND SUBTITLE

COMMUNICATIONS CHANNEL NORMALIZATION TECHNIQUES

5. FUNDING NUMBERS

C - F30602-94-C-0062
PE - 62702F
PR - 4594
TA - 15
WU - K9

6. AUTHOR(S)

Dr. Nevang Naik and Dr. Richard Mammone

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Rutgers University
Office of Research and Sponsored Programs
P.O. Box 1089
Piscataway NJ 08855-1089

8. PERFORMING ORGANIZATION REPORT NUMBER

N/A

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Rome Laboratory/IRAA
32 Hangar Rd
Rome NY 13441-4114

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

RL-TR-95-257

11. SUPPLEMENTARY NOTES

Rome Laboratory Project Engineer: John Grieco/IRAA/(315) 330-4024

12a. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT (Maximum 200 words)

Performance of Speech and Speaker recognition systems generally degrades when there is a mismatch between training and testing conditions. A significant part of this mismatch is caused by the differences in transmission channels and transducers. Performance is particularly impaired when short training and testing utterances are used. There is much interest in making systems robust to these variations. Conventional methods attempt to minimize the channel mismatch by attenuating or modifying features sensitive to channel differences.

This report describes a new methodology for extracting robust features based on systematic selection and filtering of the eigenmodes. The poles and the corresponding modes of speech are investigated under mismatched conditions caused by varying channel conditions for speaker identification systems. A method based on Pole filtering is introduced to estimate and normalize cross channel differences. Experiments on a few standard databases show improved recognition accuracy over conventional methods. In addition, Pole filtering is shown to be useful in identifying the type of channel present.

14. SUBJECT TERMS	15. NUMBER OF PAGES
Eigenmodes, Normalization	96
	16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UL

Contents

List of Figures

Abbreviations

LP is Linear Prediction
MA is Moving Average
AR is Autoregressive
FT is Fourier Transform
LPCC is Linear Predictive Cepstral Coefficients
PFCC is Pole filtered Cepstral Coefficients
CCF is Channel Compensation Filter
CMN is Cepstral Mean Normalization
CMS is Cepstral Mean Subtraction
LTM is Long Term Mean

Terms

Terms *Channel distortions* and *Convolutional distortions* have been used interchangeably.
Terms *Channel compensation* and *Channel normalization* have been used interchangeably.
Terms *Cepstral Mean Normalization* and *Cepstral Mean Subtraction* can be used interchangeably.
Term *Pole-filtering* is different from conventional filtering. It implies a filtering, weighting or a selection of poles.
Terms *True Speech* and *Clean Speech* refer to speech not subjected to any environmental degradations.

Abstract

Performance of Speech and Speaker recognition systems generally degrade when there is a mismatch between training and testing conditions. A significant part of this mismatch is caused by the differences in transmission channels and transducers. Performance is particularly impaired when short training and testing utterances are used. There is much interest in making systems robust to these variations. Conventional methods attempt to minimize the channel mismatch by attenuating or modifying features sensitive to channel differences.

Speech is usually modeled using an all-pole filter representation of the vocal tract. The poles represent the eigenmodes of the vocal tract in the time domain. Thus, a multimodal model of the vocal tract is implied. Each mode can be represented in the frequency domain as a spectral component with a constituent center frequency and a bandwidth. The components which represent the formant structure, provide sufficient information to recognize the speech sounds and the speaker under matched conditions. However, these components are adversely affected by channel distortions which deteriorate the system performance under cross channel conditions.

This report describes a new methodology for extracting robust features based on systematic selection and filtering of the eigenmodes. The poles and the corresponding modes of speech are investigated under mismatched conditions caused by varying channel conditions for speaker identification systems. A method based on Pole filtering is introduced to estimate and normalize cross channel differences. Experiments on a few standard databases show improved recognition accuracy over conventional methods. In addition, Pole filtering is shown to be useful in identifying the type of channel present.

Chapter 1

Introduction

1.1 Problem Definition

Robustness in Speech and Speaker recognition by computers has been a challenging research problem for the past several decades [1, 2, 3, 4, 5, 6, 7]. The ability to recognize the content of a spoken message is called Speech Recognition, while, identifying or verifying the person conveying the message falls in the category of Speaker Recognition. The robustness issue has been of much interest in applying speech based recognition systems to several practical applications. Applications embody numerous scenarios where voice is employed as a communication medium such as commercial transactions over telephones, air-to-tower communications, military and forensic applications. Many applications are designed to alleviate the burden on the end-user by providing a hands-free communication alternative.

It has been found that the performance accuracy of such recognition systems is reasonably high when the quality of speech being processed is clean (that is, not subjected to environmental degradations). The performance of the systems also depend on the abundance of phonetic components or sounds which embody the speech and speaker information. The issue of robustness arises when the estimates of these components are perturbed by either

- degradations in the application environment, or,
- physiological factors that affect the spoken dialog.

The spoken message may have been acquired over a noisy transmission channel with a limited bandwidth or the message may have been conveyed in the presence of substantial background noise. Variations in the quality of acquisition equipment (microphone transducers, coders, A/D) typify undesirable application environment conditions. Physiological degradations may be caused due to stress, health, dialect and mood of the speaker conveying the message. Changes in the speaker's physiological characteristics over time (also called aging) degrades the recognition accuracy. A robust speech or speaker recognition system may need to address one or all of the aforementioned degradations based on the application requirements. Figure (1.1) illustrates the factors affecting a typical recognition system.

Degradations incurred on the spoken message changes its characteristics and affects the performance of the system. Recognition systems deployed in the field (application environment), assuming a cooperative user [1], are most significantly degraded by transmission channels, varying ambient noise levels and other practical distortions caused by recording device characteristics. Locational constraints such as room reverberations, cross-talk, distance from microphone are also critical to consistent performance. Improving recognition performance of systems where the speech signal has been subjected to channel distortions

[1] a user unaffected by physiological and psychological factors such as unwillingness to communicate, stress, mood etc.

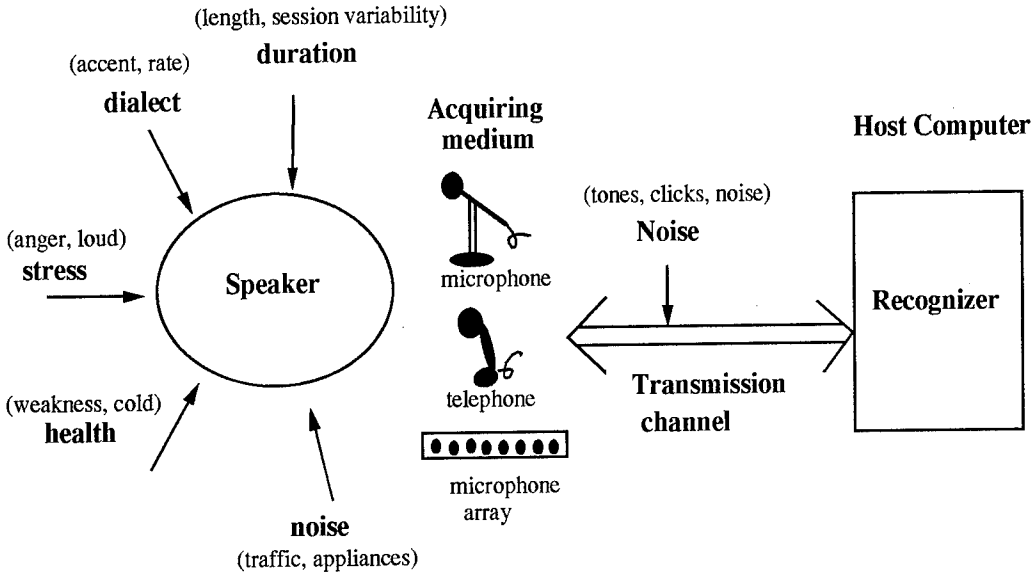


Figure 1.1: Factors influencing the performance of Recognition systems.

(due to a transmission channel, or transducer characteristics) has been of significant practical interest [26, 28, 29, 30, 32, 39, 68, 76][2] and defines the problem of research for this report.

1.2 Focus of Research

Design and development of robust speaker recognition systems has been actively addressed in the literature due to its commercial viability engendered by the rapid increase in computational power of computers [2, 9, 32, 41]. Speaker recognition applications primarily focus on security issues in commercial areas such as accessing customer banking or credit-card accounts through the use of telephonic transactions, entry of personnel in restricted areas, routing and acquiring personal messages etc. Secure access issues in military communications and forensic applications have also been addressed in the past [96, 97].

The performance of a speaker recognition system is primarily limited by availability of acquired speaker information and environmental degradations. The acquisition of a speech signal over a limited transmission bandwidth results in a substantial loss of speaker information. Speech utilized in training and testing of recognition systems acquired via different calling conditions or spoken using different handsets or microphones, causes a mismatch in the representation of a speaker's identity. Such channel mismatch causes a signification degradation in the recognition or verification of the speaker's identity. Proper channel compensation must be carried out on the speech representation in order to normalize the effects of the distortion and retain maximum speaker information contained within the available bandwidth. Developing improved methods for channel normalization (or compensation) in speaker recognition systems is the focus of research in this report.

The remaining sections in the chapter briefly review the different elements in a speaker recognition system.

[2]locational variations are usually assumed to be consistent.

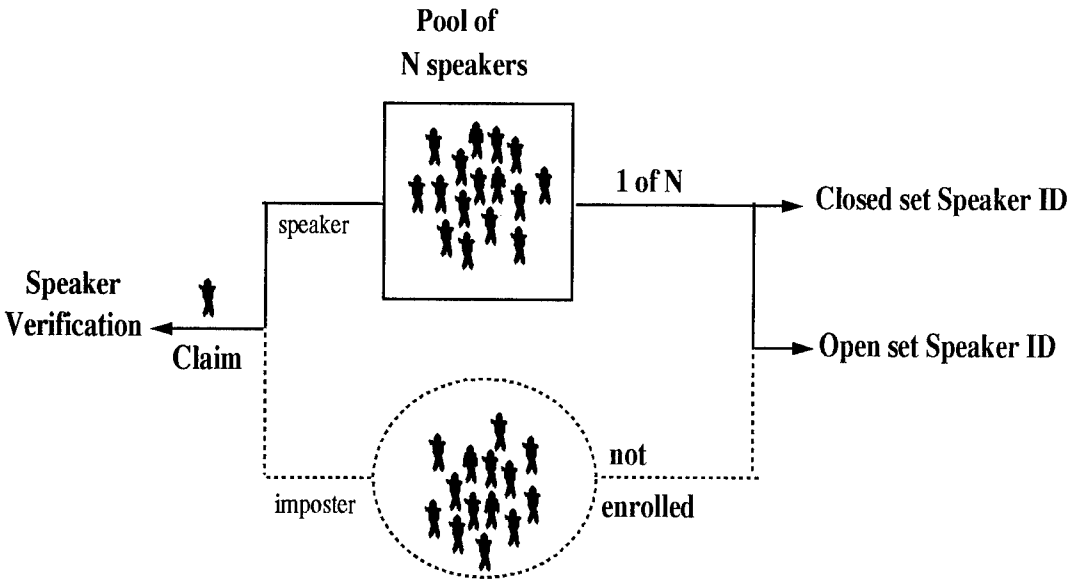


Figure 1.2: Types of speaker recognition systems.

1.3 Dichotomy of Speaker Recognition Systems

Voice recognition systems can be viewed as performing two distinct steps. The first step is *feature extraction* and the second is *classification*. The feature extraction phase involves encoding the relevant information from a speech signal in a robust manner into features or patterns. The classification phase involves modeling and matching the patterns on the basis of the observations.

A speaker recognition system may be concerned with identifying or verifying the person conveying a spoken message. Speaker recognition systems are generally classified into Speaker Identification or Speaker Verification systems [2, 3].

A Speaker Identification system determines the identity of a person from a spoken utterance, whose voice best matches one of the N voices known to the system. A speaker identification system can be *closed set* wherein the system identifies the voice of an unknown speaker that is among the N voices known to the system. An *open set* speaker identification system on the other hand, would be able to determine whether an unknown speaker belongs to the group of speakers enrolled in the system or not, and determines the speaker's identity, if enrolled [3]. A speaker verification system, *verifies* that the utterance belongs to the person who the speaker claims to be. Figure (1.2) illustrates the categories of speaker recognition systems.

In either case, the systems could be text-dependent or text-independent (or free-text). In a text-dependent system, the system is provided with a fixed phrase (decided apriori) during training and testing, as opposed to a text-independent system where the vocabulary of the spoken utterance is unconstrained.

In speaker recognition systems, both training and testing phases involve a preprocessing and feature extraction stage. In this stage, the speaker information is encoded from the speech signal into features that form a compact representation. In the training phase, the extracted features are used to build parametric or non-parametric models for each speaker in the recognition system. The testing phase involves extracting features and finding a best match to the existing speaker models built during training. Such pattern matching is carried out using a classifier. Figure (1.3) illustrates the basic elements of a speaker identification system and Figure (1.4) illustrates a typical speaker verification system over a telephone link.

Practical speaker recognition requires robust feature extraction methods that demonstrate an **invariance** to uncontrollable application environment parameters. Robust classification schemes are also required

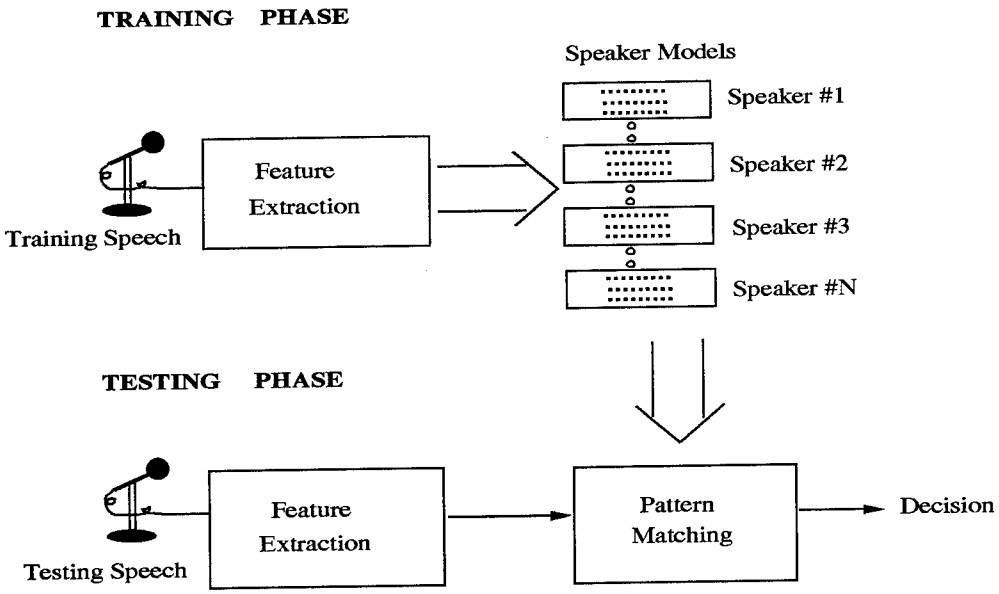


Figure 1.3: A typical speaker identification system.

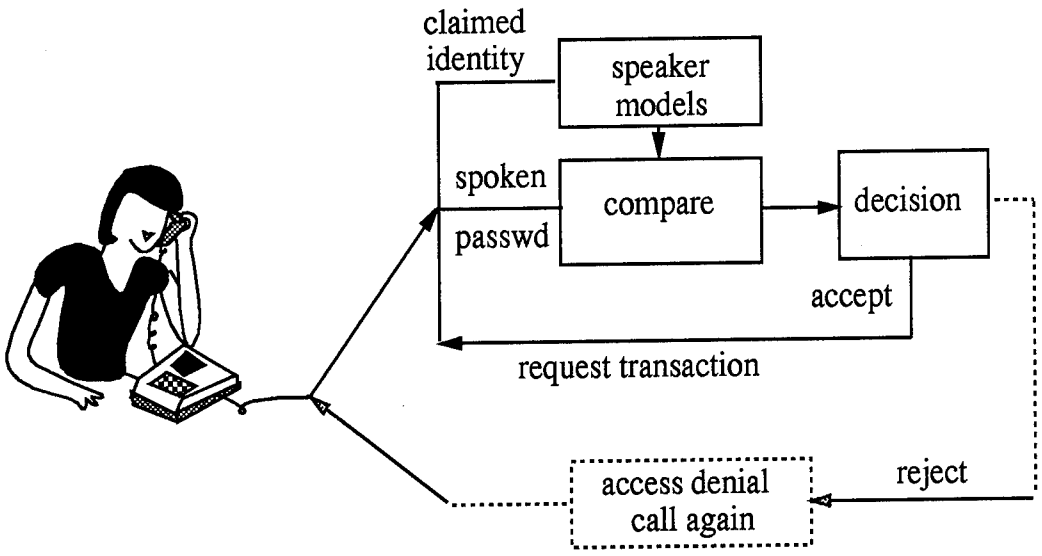


Figure 1.4: A typical speaker verification system (adapted from Rabiner [9]).

that best model the characteristics of speech unique to an individual in order to maximize the discriminability among speakers, independent of the population. The overall recognition system may be expected to exhibit a graceful degradation in the accuracy with increased speaker population.

Besides channel distortions and additive noise, limited availability of data for training and testing in practical speaker recognition systems is a critical factor in their viability. Insufficient availability of speech data negatively affects the quality of the features and results in poor estimates of distortions. Insufficient training data also imposes a modeling constraint on the richness of the classifiers. Classifiers that offer improved performance without the need for substantial training data for modeling are desirable for practical speaker recognition.

The issue of robust speaker recognition can be addressed by categorizing it into modeling stages that involve preprocessing and feature extraction followed by optimized training and classification. Past literature is replete with methods that have focused on all stages of speaker recognition. Comprehensive reviews have been published in references [1, 2, 3, 4, 24, 41]. Robust feature extraction has predominantly focused on the issues of channel degradations and ambient noise.

In this report a robust feature extraction methodology called **Pole Filtering** is introduced to improve the performance of conventional speaker recognition systems degraded by convolutional distortions [3]. A significant part of the report focuses on using a conventional signal modeling technique called Linear Predictive (LP) modeling for spectral analysis of speech. The eigenmodes of a linear system modeling a segment of speech and its relation to LP analysis is investigated. The study of eigenmodes of speech and their perturbations to convolutional distortions forms the basis of the contribution. The new technique constitutes a **filtering**, **weighting** and **selecting** of the eigenmodes of speech so as to reduce the effect of channel distortions.

The report is organized as follows: Chapter two provides a review of conventional speaker recognition. Chapter three reviews methods for robust features for channel normalization. Chapter four introduces the philosophy of Pole-filtering for robust feature extraction in speaker recognition, followed by experimental results on various benchmark databases in Chapter Five. Chapter Six introduces an upshot of the pole-filtering methodology an approach to channel identification based on speech. Chapter seven gives the summary and conclusions of our findings along with a perspective on future work to consolidate the methodology.

[3] consisting of an overall convolutional distortion due to the transmission channel and the microphone or handset.

Chapter 2

Conventional Speaker Recognition

The process of speaker recognition and the underlying issue of robustness can be investigated by subdividing it into

- **Front-end modeling** for preprocessing and feature extraction, and,
- **Back-end modeling** for Classification,

as shown in Figure (2.1). The recognition process can also be unified under a signal modeling paradigm, where the speech signal is first parameterized into a "perceptually meaningful" representation and then modeled statistically for robust classification.

The speech signal is usually acquired digitally through an A/D conversion process, the purpose of which is to produce a digital representation of the speech signal with as high a signal-to-noise ratio (SNR) as possible.

The non-stationary characteristics of a speech signal necessitates that the signal be parameterized in time slices (frames) short enough to assume stationarity so that conventional signal modeling techniques can be applied. For robust recognition, parameterizations are sought that are invariant to transmission channel variations, transducer characteristics and ambient noise. Parameters from a signal, also called *feature vectors* or *observations* are then modeled and classified. The subsequent sections elaborate on each stage of the speaker recognition process.

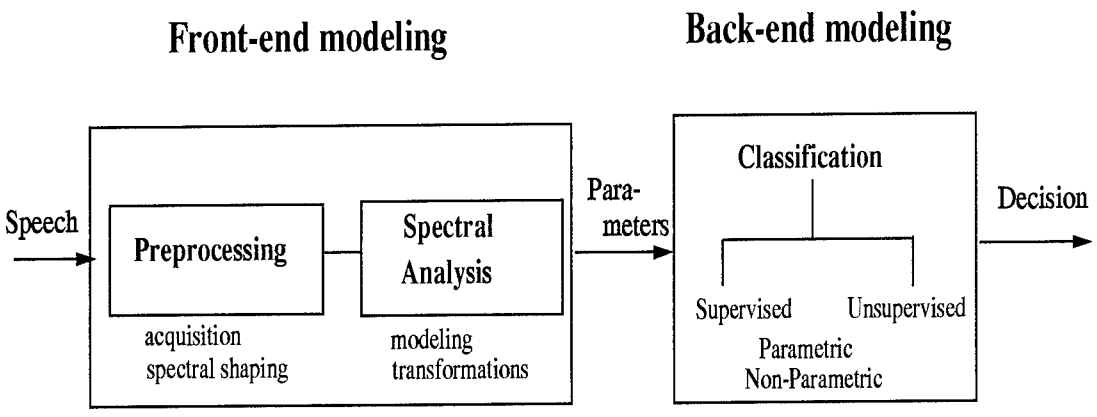


Figure 2.1: Speaker recognition process.

2.1 Conventional Front-end Modeling

Front-ends for speaker recognition typically consist of a preprocessing step followed by a feature extraction step.

2.1.1 Preprocessing

A short-time analysis of the speech signal is carried out by segmentation into overlapping frames which are typically 15-30 msec, with an overlap of 10-15 msec. Within this duration the properties of a speech signal can be assumed to be stationary.

Speech analysis is effectively carried out in the spectral domain since speech signal parameters are found to be more consistent in the spectral domain, and information from the spectral domain is more easily encoded. Thus, speaker recognition systems have generally employed features obtained via processing in the spectral domain. The preprocessing step involves spectral shaping of the speech signal to emphasize important frequency components in the speech signal. Voiced sections of a speech signal, which are predominantly used for speaker recognition, naturally have an attenuation of approximately 20 dB per decade due to the physiological characteristics of the speech production system [13]. The spectral shaping is done with the use of a *Preemphasis* filter that offsets the negative spectral slope thereby improving the analysis of speech [13, 14]. A single tap preemphasis filter,

$$H_{pre}(z) = 1 + a_{pre}z^{-1}, \tag{2.1}$$

where, $a_{pre} \in [-1.0, -0.4]$ [1] is normally employed.

For speaker recognition, preemphasis has an effect of enhancing the speaker information in the higher frequency bands of the spectrum. However, preemphasis is helpful only in modest signal-to-noise ratios since noise affects the estimates of higher frequency components adversely. Preemphasis often degrades the performance of speaker recognition systems under low SNRs.

2.1.2 Feature extraction based on Spectral Analysis

Spectral analysis is used to extract features by encoding the speech waveform into meaningful parameters that assist classification. Spectral analysis retains only those components in the spectral representation of the speech frame that are sufficient for modeling and recognition purposes.

Two spectral analysis methods have been widely employed based on the application domain. The methods are either based on

- Fourier transform (FT) modeling, or,
- Linear Prediction (LP) modeling.

Parameterizations are usually derived from these modeling techniques using either Filter bank analysis or Cepstral analysis as shown in Figure (2.2). The filter bank analysis technique has been improved by spacing the banks along a perceptual frequency scale to analyze the information content of speech in the different subbands. Many perceptual spacing techniques for the filters have been proposed [69, 70]. The energy within each subband yields an encoded representation of the spectra. An equally spaced Q–channel filter bank has been shown in Figure (2.3).

Modeling using Fourier transforms computes the filter bank amplitudes using a DFT (Discrete Fourier Transform) or an FFT (Fast Fourier Transform) by simply evaluating the spectrum at a discrete set of frequencies. Filter bank amplitudes for the Linear Prediction model are derived by sampling the LP

[1] A typical value of -0.95 is frequently used.

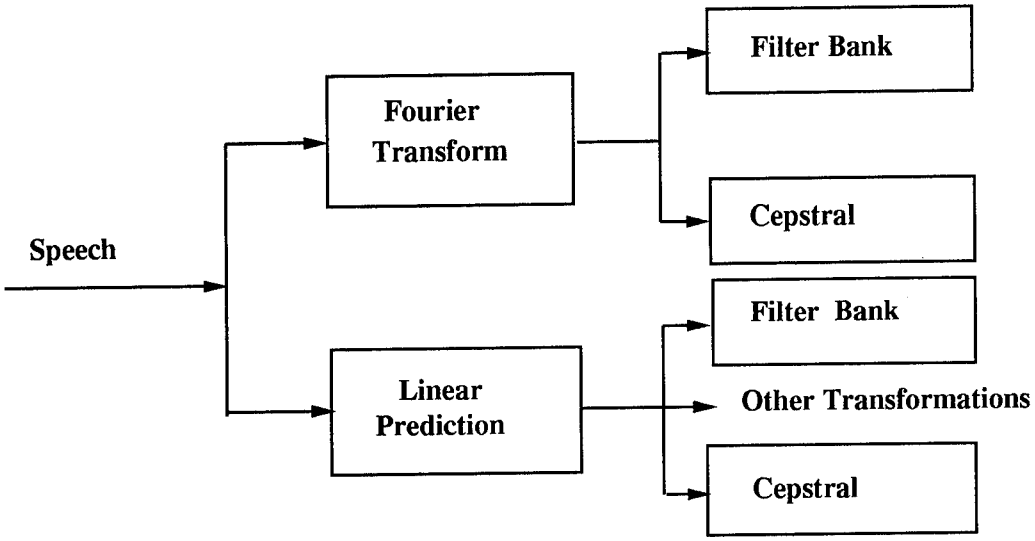


Figure 2.2: Spectral analysis methods.

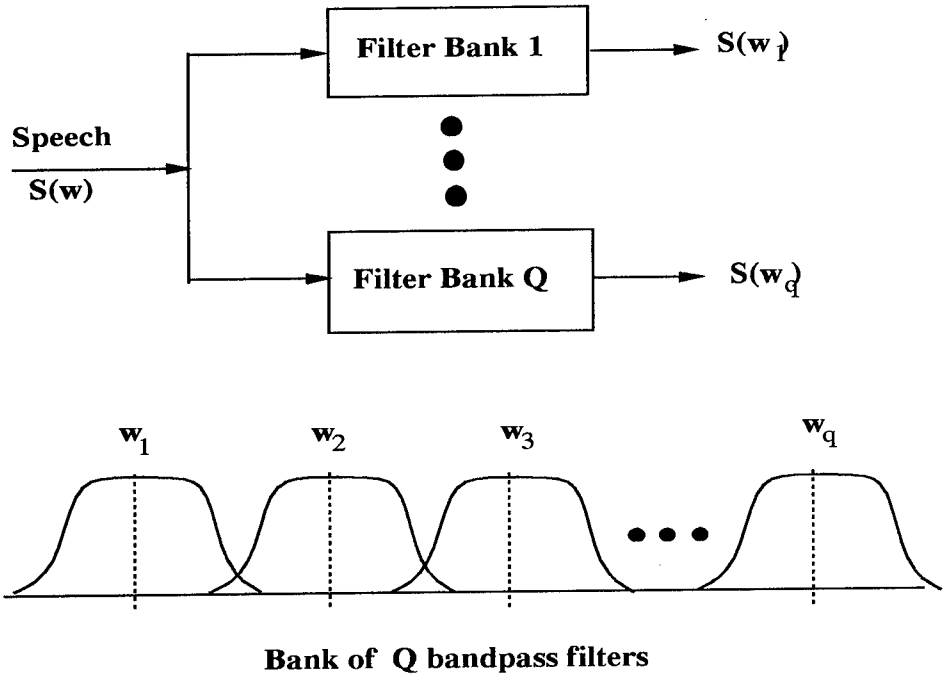


Figure 2.3: Q-channel equally spaced filter bank.

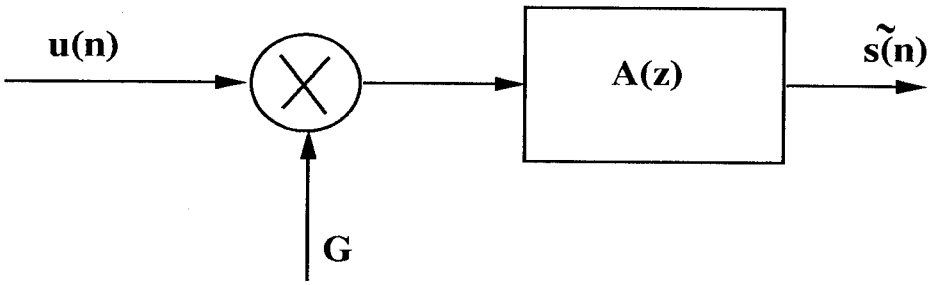


Figure 2.4: LP model of speech.

spectrum instead of the signal spectrum at the appropriate filter bank frequencies with the desired nonlinear warping. Another way involves computing the power spectrum from the autocorrelation of the impulse response of the LP filter [7].

For speaker recognition, a parametric transformation called the **Cepstral** transformation has been proven to be the most robust for extracting features [1, 7, 32]. LP derived cepstral transformation in particular has been found to provide robust statistics for speaker recognition. Filter bank amplitudes derived from the Fourier Transform based modeling are transformed into the cepstral domain via a DCT (Discrete Cosine Transform). Cepstral analysis based on linear prediction is the focus of investigation in this report. The cepstral transformation has been proven to exhibit invariance to transmission channel distortions [29, 32], an important property that shall be exploited in this report for extracting robust features. The subsequent sections outline the Linear Prediction modeling and cepstral analysis which form the nucleus of the Pole-filtering methodology.

2.1.3 Linear Prediction based modeling

Autoregressive (AR) models have been employed to model the speech production process for some time [1, 13]. The speech signal primarily consists of voiced and unvoiced sounds that can be modeled using an AR process. Thus, speech production can be viewed as an acoustic filtering operation in which an acoustic source excites a vocal tract filter [14, 17]. The effect of the excitation (glottal) and the vocal tract (acoustic filter) can be represented by a time-varying (all-pole) digital filter whose transfer function is given by,

$$H(z) = \frac{S(z)}{U(z)} = \frac{G}{1 - \sum_{k=1}^{P} a_k z^{-k}}, \tag{2.2}$$

where, U(z) is the glottal excitation transfer function corresponding to an excitation $u(n)$, G is the gain parameter, a_k are the vocal tract filter coefficients and p is the order of analysis.

The Linear Predictive (LP) technique provides a tool for parameterizing the all-pole vocal tract filter model. The technique attempts to optimally model the spectrum of a segment of speech as an autoregressive process [13]. Given a speech signal $s(n)$, the signal is modeled as a linear combination of its previous samples. The sampled signal $s(n)$ and the excitation $u(n)$ are related by the following difference equation,

$$s(n) = \sum_{k=1}^{P} a_k s(n-k) + Gu(n). \tag{2.3}$$

The interpretation of equation [2.3] is given in Figure (2.4).

A linear predictor with prediction coefficients, α_k, is defined by a system whose output is

$$\widetilde{s(n)} = \sum_{k=1}^{P} \alpha_k s(n-k). \tag{2.4}$$

The prediction error can then be defined as

$$e(n) = s(n) - \tilde{s(n)} = s(n) - \sum_{k=1}^{P} \alpha_k s(n-k). \tag{2.5}$$

The error sequence can be interpreted as an output of a system with a transfer function

$$A(z) = 1 - \sum_{k=1}^{P} \alpha_k z^{-k}. \tag{2.6}$$

If $\alpha_k = a_k$, and the speech signal obeys the difference equation, then $e(n) = Gu(n)$. This allows the prediction error filter to be an inverse filter for the system, $S(z)$, [14],

$$S(z) = \frac{G}{A(z)}. \tag{2.7}$$

The equation [2.5] may also be rewritten in Z-transform notation as a linear filtering operation,

$$E(z) = A(z)S(z). \tag{2.8}$$

The basic goal of linear predictive analysis is to solve for the set of predictor coefficients. The inverse filter, $A(z)$, provides a good estimate of the spectral properties of the speech signal. In fact the filter $A(z)$, effectively models the short-time spectrum of the signal as a smooth spectrum [13]. Since the spectral properties of speech vary over time, the predictor coefficients at a given time need to be estimated from a short windowed segment of the speech signal occurring around that time. The basic approach finds the set of prediction coefficients that minimizes the mean square prediction error over the short windowed segment of the speech waveform. There are two common methods to solve for the predictor coefficients: the *Autocorrelation analysis* and *Covariance analysis* methods. The LP filter parameters derived using autocorrelation analysis guarantee a stable (minimum-phase) filter. Details of the methods can be obtained in standard references on spectral analysis [13, 11].

The predictor coefficients represent a normalized spectrum independent of the power of the signal. In order for the spectrum from the LP model to match the spectrum of the original signal, the gain term has to be evaluated. The gain term is usually ignored for recognition purposes to allow the parameterization to be independent of the signal intensity. The gain term, however, is essential for applications such as speech coding and synthesis. The roots of the LP inverse filter correspond to the poles of the filter. Predictor coefficients or the derived poles of the filter and their transformations have often be used for recognition purposes. The transformations include [13],

- LP filter coefficients
- LP cepstral coefficients (LPCC)
- Line spectral frequencies
- Pseudo Vocal tract area functions

Atal [29] provided a comprehensive review and comparison of parameters based on LP analysis like autocorrelation, pseudo vocal tract area function, fundamental frequency and LP cepstral coefficients for the purpose of speaker recognition. The following section reviews the cepstral domain analysis of speech and its inherent property of robustness against channel distortions. There exists an important relationship between the LP parameters and cepstral parameters which shall be established in the subsequent chapters.

2.1.4 Homomorphic processing and Cepstral analysis

The problem of deconvolving and separating a signal $s(t)$ convolved with another signal $h(t)$ appears in many contexts in signal processing such as image restoration [102, 103], echo removal [101], communication [104, 105] etc. In the case when both the signals are unknown, the problem of estimating and eliminating one of the unknown signals is referred to as *Blind Deconvolution* [99, 100, 106]. The problem of blind deconvolution of the convolutional distortion for a speech signal transmitted over a telephone line and/or acquired via a transducer occurs frequently in speech processing.[2]

For speech signals the channel distortion is assumed to be time-invariant or varying much more slowly than the variations in the signal. Due to the non-stationary nature of speech signals, channel estimation is carried out on speech segments over which the signal can be assumed to be stationary.

Homomorphic systems are a class of nonlinear systems that obey the generalized principle of superposition for convolution [18]. The homomorphic theory maps the process of convolution into one of addition which simplifies the deconvolution process into a subtraction.

For a voiced frame of a speech signal, let,

$$Y(\omega) = S(\omega)H(\omega), \tag{2.9}$$

where, $S(\omega)$ corresponds to the frequency response of speech and $H(\omega)$ is the response of the channel. In time domain, $H(\omega)$ corresponds to $\mathbf{h}$, the impulse response of the convolutional distortion.

Taking the logarithm (complex), on both sides,

$$log(Y(\omega)) = log(S(\omega)) + log(H(\omega)). \tag{2.10}$$

The complex cepstrum of a signal is defined as the Fourier transform of the log of the signal spectrum. Then, for a magnitude-square spectrum $|Y(\omega)|$, which is symmetric with respect to $\omega = 0$, the Fourier series representation of $log(|Y(\omega)|)$ can be expressed as,

$$log(|Y(\omega)|) = \sum_{n=-\infty}^{n=\infty} c_n e^{-j\omega n}, \tag{2.11}$$

where $c_n = c_{-n}$ and c_n are referred to as the **cepstral coefficients**.

For a stable (minimum phase) all-pole filter modeling of speech which is generally obtained using the autocorrelation analysis method for LP modeling,

$$|Y(\omega)| \Rightarrow \frac{\sigma^2}{|A(e^{j\omega})|^2}, \tag{2.12}$$

and it is possible to define the cepstral coefficients as,

$$log(\frac{\sigma^2}{|A(e^{j\omega})|^2}) = \sum_{n==\infty}^{n=\infty} c_n e^{-jn\omega}. \tag{2.13}$$

[2]In practice it would be impossible to estimate the individual distortions caused by a transmission channel or the transducer. The term *convolutional distortion* implies an overall convolutional effect of the distortions.

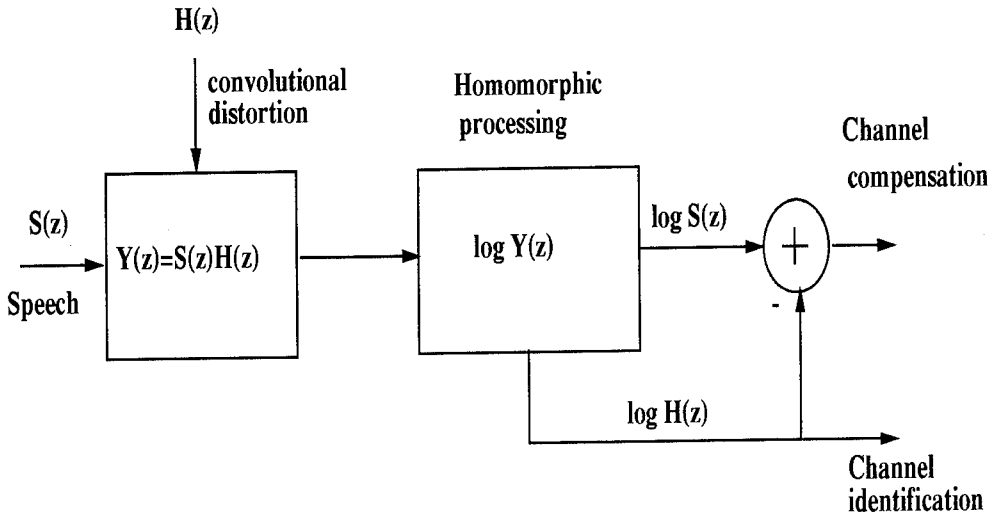


Figure 2.5: Homomorphic processing approach.

where, the cepstrum derived from the minimum-phase all-pole power spectrum is referred to as the **LP cepstrum**.

In either case, when investigating the cepstral coefficients of speech from a speech signal degraded by a channel, one can observe that the invariant distortion due to the channel appears as an additive component in equation [2.10].

In the case of LP derived cepstrum evaluated for M frames of the speech utterance, one can formulate equation [2.10] as,

$$\sum_{m=0}^{M-1} log(|Y(\omega;m)|) = \sum_{m=0}^{M-1} log(|S(\omega;m)|) + log(H(\omega)). \tag{2.14}$$

A time invariant distortion $H(\omega)$, can be eliminated by averaging in the cepstral domain, and then subtracting this average component. Such a *cepstral bias* (corresponding to $log(H\omega)$) can be used to eliminate channel distortions as well as estimate or detect a channel as shown in Figure (2.5).

However, this method of eliminating the time-invariant convolutional distortion relies on the broad assumption that

$$\sum_{m=0}^{M-1} log(|S(\omega;m)|) \rightarrow 0. \tag{2.15}$$

Equation [2.15] implies that the average component due to actual speech is zero-mean.

Hence, the subtraction of the average component in the cepstral domain would essentially correspond to a deconvolution term or a channel normalization term. This term also yields an estimate of the freqeuncy response of the degradation. Since for speech systems the convolutional distortion is not known apriori, it implies a method of performing a blind deconvolution in the cepstral domain.

For speech signals, the emphasis is only on the log magnitude spectra or the log amplitudes. An accurate estimate of the channel can only be obtained if equation [2.15] is satisfied for large M. However, in practice, the number of frames available for processing a speech utterance is always limited by the application and the theoretical consideration in equation [2.15] is never satisfied. The amount of speech for training or testing is invariably too limited to yield accurate estimates of $log(H(\omega))$. Hence, the average term on the left-hand side of equation [2.10] reflects the presence of,

- **a gross spectral distribution of the phonetic content of the spoken material and speaker characteristics and**
- **a spectral distribution of the channel.**

This report focuses on a methodology to *decouple* the two components, the speech information and the channel information, to obtain an improved cepstral estimate of the underlying distortion corresponding to $H(\omega)$. LP derived cepstral features have been considered as tools to develop improved methods for channel compensation. The relation between the poles of the all-pole LP filter and their corresponding transformation to the cepstral domain is studied under convolutional distortions. A **Pole filtering** approach is developed to normalize the effects of a channel. The approach uses intelligent filtering of the pole parameters of the all-pole LP filter, that correspond to the eigenmodes of speech, in order to yield a better estimate of the channel in the cepstral domain.

For speech based recognition systems, several conventional techniques have been developed in the cepstral domain (derived via FFT or LP techniques) to compensate for the convolutional degradations [29, 30, 32, 45, 68]. Cepstral parameters are usually derived after modeling speech using LP analysis using a recursive relationship between the LP prediction coefficients and cepstral coefficients [29, 87].

Features based on LP or FFT derived cepstra have been shown to yield the best speaker recognition results. In recent speaker recognition literature, the use of cepstral coefficients as features has been dominant [2, 28, 29, 34, 36, 37, 39]. Characteristics of the cepstral coefficients have been extensively studied for minimizing mismatch caused between training and testing features due to the transmission channel. Processing has been emphasized in two domains: the *Intraframe domain* (processing within each speech frame) [30, 72, 73], and the *Interframe domain* (processing across an ensemble of speech frames) [29, 32, 68].

The subsequent chapter provides a comprehensive overview of the approaches along with conventional features for channel normalization. The remaining sections in this chapter survey back-end modeling or classification schemes that have been employed for speaker recognition.

2.2 Back-end Modeling or Classification

Robust classification is concerned with the decision making process in determining or verifying the identity of the speaker of a spoken utterance. In general, parameterizations of the training utterance of a speaker are modeled to form a reference template. The test utterance is then parameterized and a pattern matching algorithm is used to determine which reference template best matches the test utterance.

The parameters derived from an utterance normally represent a set of points in a multidimensional parameter space. A statistical average of the features may be used to distinguish the speakers [53, 33, 54]. The simplest template matching technique compares the test utterance to the training template by computing the distance between the feature means. A Euclidean distance is used for minimum distance classification given by

$$(\bar{x} - \mu)'(\bar{x} - \mu). \tag{2.16}$$

where μ is the mean that corresponds to a reference parameters and $\bar{x}$ is the mean of the test parameters. The superscript $'$ denotes a transpose. If the mean and the covariance Σ are known then a weighted Euclidean distance measure also known as the Mahanalobis distance given by,

$$(\bar{x} - \mu)'\Sigma^{-1}(\bar{x} - \mu), \tag{2.17}$$

and the estimated covariance $(\tilde{\Sigma})$ is given by,

$$\tilde{\Sigma} = \frac{1}{n}\sum_{i=1}^{M}(x_i - \bar{x})(x_i - \bar{x})', \tag{2.18}$$

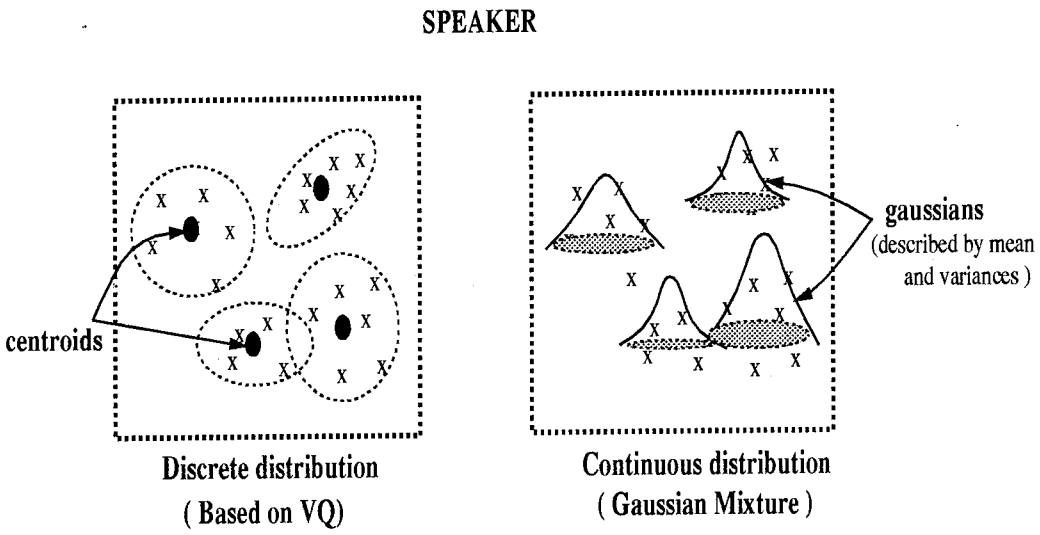


Figure 2.6: Parametric v/s non-parametric modeling.

where x_i corresponds to the feature for the i^{th} frame. The long term average of feature vector is however sensitive to variations in the background noise and channel distortions.

More sophisticated methods of analyzing speaker information in feature space have been investigated in the past [8, 34, 35, 24, 40]. The classification methods can be broadly characterized from their training methodologies. The speaker modeling for classification can either be *Supervised* or, *Unsupervised.*

In supervised training methods, the classifier is trained by associating the feature data to a category or a class label. In case of speaker recognition systems, each speaker would correspond to a category. In unsupervised training methods, the classifier is allowed to model each speaker usually based on a clustering of speaker-dependent sounds from the spoken utterance or utterances. During the testing phase, the speaker is classified as being the speaker which is most likely to have generated those sounds, by matching the test patterns to all existing speaker models.

Classifiers that have been designed based on Supervised and Unsupervised training for speaker recognition are reviewed in the subsequent sections.

2.2.1 Unsupervised Classifiers

For unsupervised classification, speaker information is modeled statistically by imposing a model on the data [1]. Generally, a multivariate Gaussian density function is chosen. The modeling could be *parametric*, where a continuous underlying distribution is assumed or *non-parametric*, wherein a discrete distribution is assumed. The Gaussian density function has the property that it is completely characterized by its mean vector and the covariance matrix. Most classifiers assume this distribution when modeling the speaker.

For some classifiers, parametric fits that are based on gaussian statistics may not be appropriate since abundant data may be required to model the underlying statistics of the speaker. In case of limited training data, a non-parametric fit is often performed by hypothesizing a discrete probability distribution. Figure (2.6) illustrates the difference between parametric and non-parametric fits to the data. One such method of realizing a non-parametric fit is used in the *Vector Quantizer* (VQ) classifier [84, 85].

Vector Quantization is an unsupervised classification method. The feature parameters from the speaker's utterance are grouped or clustered into data representatives that form a compressed representation of the speaker's feature space. All vectors falling inside a cluster are represented by a centroid or a local statistical mean. Thus the feature space of the speaker is quantized to a *Codebook* of centroids called *Codewords.*

Heuristically, each codeword may be thought of modeling speech components, such as speaker dependent phonemes or sounds.

Several clustering algorithms exist in the literature [84, 34, 11, 15, 85]. The techniques involve selecting an initial estimate of the codewords and then iteratively updating the centroids such that the average distance of vectors from their nearest centroid is minimized. Unsupervised k-means is a commonly used algorithm for building codebooks representing the speaker model [84]. During the testing phase, an accumulated distance measure is used to match the test patterns to each speaker's codebook. The speaker that corresponds to a minimum accumulated distance is selected.

Several parametric models based on unsupervised training have been proposed wherein a likelihood approach is used to classify the speaker. The distribution for a speaker is assumed to have continuous densities p_i and a likelihood $p_i(x)$ is associated with a feature x, generated by the i^{th} speaker. Using Bayes' theorem, the probability that the speaker is the i^{th} speaker is given by,

$$P(speaker = i|x) = \frac{p_i(x)P_i}{p(x)}, \tag{2.19}$$

where P_i is the *apriori* probability that the utterance was spoken by speaker i, and $p(x)$ is the probability of the feature x occurring from any speaker. If the prior probabilities are equal and $p(x)$, which is the average of the speaker densities, is the same, the classified speaker will be chosen as the speaker that has the highest likelihood score.

Given, n independent feature vectors, from a speaker's utterance, $X = x_1, x_2, \cdots, x_n$, for a gaussian model with parameters μ and Σ, the likelihood of the test utterance is given by [41],

$$L(X; \mu; \Sigma) = |2\pi\Sigma|^{-\frac{n}{2}} e^{-\frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)\Sigma^{-1}(x_i-\mu)'}. \tag{2.20}$$

For computational convenience, log-likelihoods are generally calculated. Gaussian models attempt to model the gross distribution of the speaker. In recent literature, Gaussian Mixture Models (GMM) [37, 41, 58] have been shown to offer a viable robust speaker recognition model. GMMs model the speaker information as a mixture of Gaussian distributed data clusters with weighted sum of densities. The mixture model is trained by partitioning the frames of an utterance into a predetermined number of clusters. This is usually carried out with a clustering algorithm or by automatic speech segmentation. The training is then carried out with an Estimation-Maximization algorithm[95].

In recent literature, Hidden Markov Models (HMMs) have also been used for speaker recognition[57, 59, 60, 56]. HMMs comprise of employing a stochastic finite state machine to model sequences. HMMs employ a markov chain consisting of a sequence of states. For each state, a posteriori probability is computed along with a transitional probability from state to state. HMMs attempt to model the state transitional probabilities of a specific speaker. In fact, an HMM modeled with a discrete distribution may be considered as a Vector Quantizer that additionally utilizes transitional information from state to state. A comprehensive comparison of VQ and HMM classifiers was reviewed by Matsui and Furui wherein it was concluded that an HMM is as robust as VQ for a text-independent speaker identification task (which shall be focussed in this report) when enough data was available [56]. Unsupervised classifiers have a drawback in that they do not have discriminatory information. Such discriminatory information is particularly useful in case of short training and test utterances.

2.2.2 Supervised Classifiers

Supervised classifiers involve training a classifier for a speaker with the knowledge of all other speakers as a "pool of anti-speakers" or "not-the-speaker". In general, individual classifier models are trained for all the speakers. Each classifier is trained with the feature vectors from a speaker labeled as "ones" and the feature

vectors for the remaining speakers labeled as "zeros". The classifier is trained with all these labeled feature vectors. A classifier for each speaker in the population is trained using this method. During testing, the test vectors for a specific speaker should yield a "one" response for that speaker's classifier and a response of "zeros" for the feature vectors of all other speakers or "anti-speakers". For speaker identification, the test vectors are applied to each speaker's classifier and the speaker is selected as corresponding to the classifier with the maximum accumulated output. Supervised classifiers such as Neural tree Network (NTN) [86], Multi-layer Perceptrons (MLP) [21], Time Delay Neural Networks (TDNN) [78] and Radial Basis Function (RBF) [98] have been successfully employed for speaker recognition [24, 48, 61, 62, 63].

While MLP and RBF classifiers require a predetermined architecture, NTNs [86] are flexible as they impose no architectural constraints since the architecture is determined while training. An NTN is a hierarchical classifier that combines the properties of feed-forward neural networks within a decision tree structure. For speaker recognition, a modified neural tree network (MNTN) [24] has been used wherein, a binary NTN (dual decision) is grown for each speaker with speaker labels and a common anti-speaker labeling. All training vectors are applied to the MNTN and the tree is recursively grown until a classification criterion is satisfied [24, 48].

Speaker recognition in NTNs is carried out by recording the labels for all the test vectors. The corresponding speaker likelihood is computed as,

$$P_{ntn}(x|S_j) = \frac{M}{N+M}, \tag{2.21}$$

where M is the number of vectors classified as the speaker, N is the number of vectors classified as "anti-speaker". Figure (2.7) [24] illustrates the recursive tessellation process in the speaker and anti-speaker's feature space.

For the MNTN, the likelihood score is weighted by a confidence measure. In either case, the speaker corresponds to the model most likely to have generated the observed sequence of vectors based on an accumulated likelihood. NTN/MNTNs have fast retrieval times and are efficient for hardware implementations of speaker recognition systems. Recently, a fusion of classifiers has been proven to yield improved results over using individual classifiers. Farrell [49] considers the fusion of NTNs and VQ for speaker identification. The fusion approach exploits the advantages of VQ, which determines how close the test utterance is to the speaker's model, and NTN, which determines how different the test utterance is of the speaker from that of other speakers in the population. Soong and Rosenberg [35] used fusion of separate VQ codebooks formed using static and dynamic features of speech.

With an emphasis on short training and testing utterances, non-parametric modeling techniques, particularly VQ, have prevailed as an unsupervised classifier due to its simplicity and robustness, while the NTNs have prevailed as a supervised classification scheme, due to its fast retrieval time and discriminatory capacity. In this report, the VQ classifier is employed with empirically fixed training parameters while the features are evaluated for comparison.

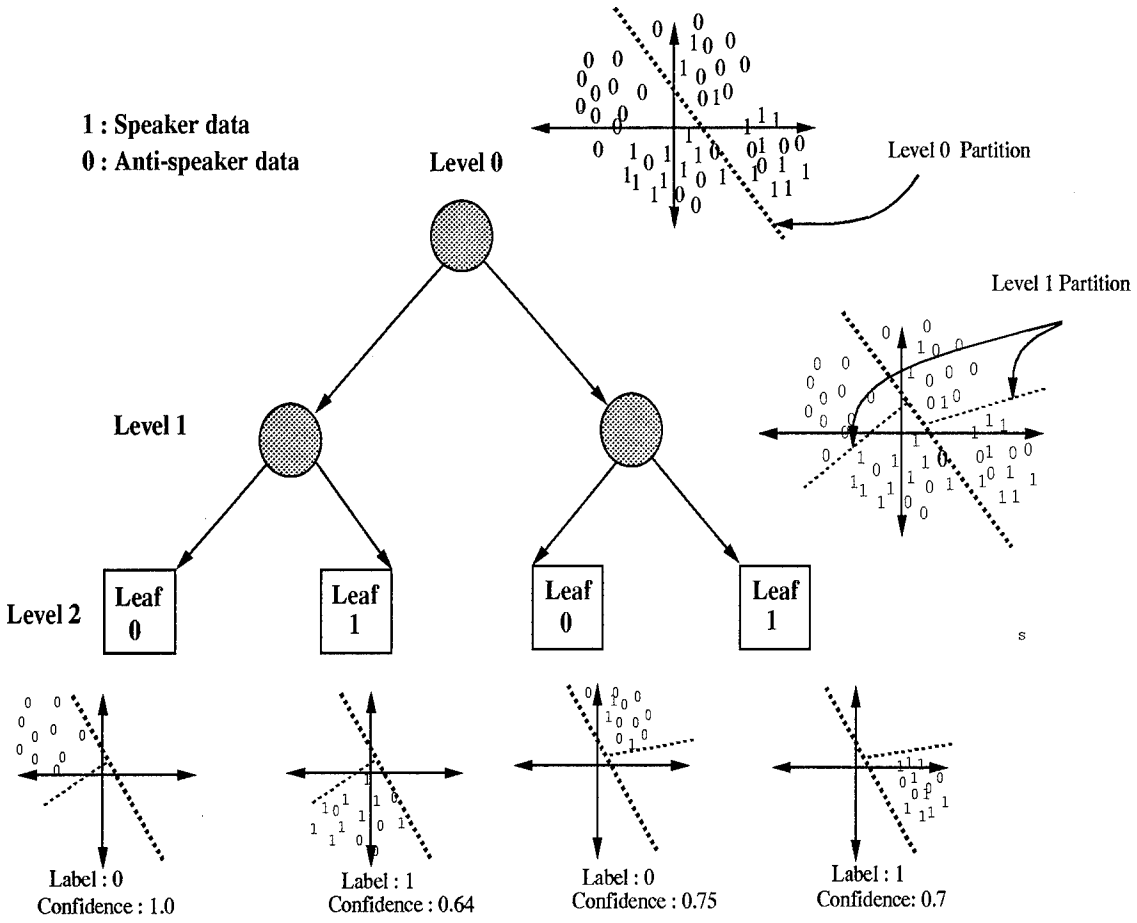


Figure 2.7: Neural Tree Network for speaker recognition (Adapted from Farrell and Mammone [24]).

Chapter 3

Robust Features for Channel Normalization

For speaker recognition systems to be practical, they must be made robust to convolutional distortions. There is a need to acquire features invariant to significantly mismatched conditions caused by:

- transmission channels (local, long-distance, radio-links). Figure (3.1) illustrates frequency responses of typical transmission channels obtained from a simulator [88],
- transducer equipment such as microphone used for recording, handsets (electret or carbon-button for telephone applications) and,
- usage condition such as a speaker phone or cellular phone.

The degradations due to channel differences incurred by the above distortions is the most critical limitation of state-of-the-art speaker/speech recognition systems [9, 41]. In practice, all these degradations are assumed to correspond to an overall convolutional distortion component that degrades the spoken utterance to be trained or tested.

Current research in channel normalization has attempted to improve recognition accuracies of speech and speaker recognition systems by either

1. extracting features invariant to channel mismatch, or,
2. modeling and estimating the composite effect of the convolution in a probabilistic framework.

This chapter outlines the conventional approaches to channel normalization with respect to extracting features invariant to mismatch caused by the convolutional distortions. It will be proven that all conventional approaches share a common underlying basis for channel normalization, which in itself has a drawback. The investigation of the inherent limitations of these approaches shall be used as a basis for developing improved methods for normalizing the convolutional differences. Probabilistic channel modeling and estimation will be reviewed briefly towards the end of the chapter and is not focussed in this report.

3.1 Features for Channel Normalization

Spectral modeling of practical speech signals require modeling the channel and noise distortions in addition to sub-glottal effects, vocal tract resonances and other physiological characteristics. A proper spectral modeling technique would ideally need to model each of the effects with their respective transfer functions. A pole-zero approximation of the transfer functions consists of an Autoregressive (AR) component and

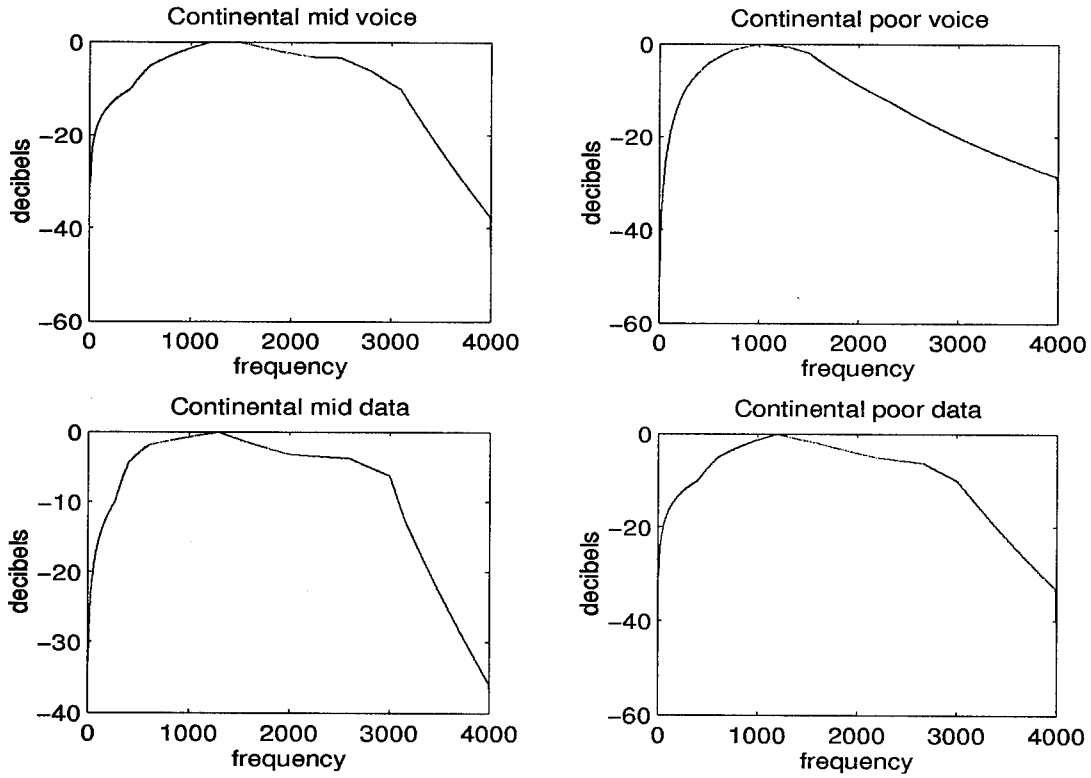


Figure 3.1: Various transmission channel characteristics.

a Moving Average (MA) component. In general, these factors can be represented by individual transfer functions consisting of poles and zeros multiplicative to the transfer function of speech.

In practice, a transfer function for speech that takes into consideration all the factors can be represented by,

$$S(z) = \frac{N_{nasals}(z)N_{channel}(z)N_{noise}(z)}{A_{resonances}(z)D_{channel}(z)} + \frac{N_{add}(z)}{D_{add}(z)}, \tag{3.1}$$

where $N_{add}(z)$ and $D_{add}(z)$ model additive noise. $N_{channel}(z)$ and $D_{channel}(z)$ represents the zeros and poles of the channel distortion. $N_{nasal}(z)$ models the nasal zeros in the speech spectra.

ARMA modeling of speech has been found to be very complicated and computationally exhaustive in practice. Optimization techniques based on maximum likelihood estimation (MLE) and related concepts have often been used to determine the AR and MA parameters [22]. Many of these optimal and sub-optimal techniques are unreliable and have convergence problems. For recognition purposes, AR modeling of the speech signal has been found to suffice within modeling constraints such as filter order and numeric precision [14].

An autoregressive fit to the spectrum of a segment of speech, however, corresponds to an all-pole approximation that subsumes all environmental factors affecting speech. Linear Prediction modeling yields an all-pole fit to the spectra representing all the above convolutional and additive factors in equation [3.1]. A robust parameterization based on LP analysis would ideally require a modeling and compensation for the individual distortions and a methodology to decouple the spectral information related to only the speech or the speaker.

Parametrizations based on AR modeling are particularly significant when derived for speech signals employed in practice. Although telecommunication systems today, regularly deliver SNR's in excess of 30

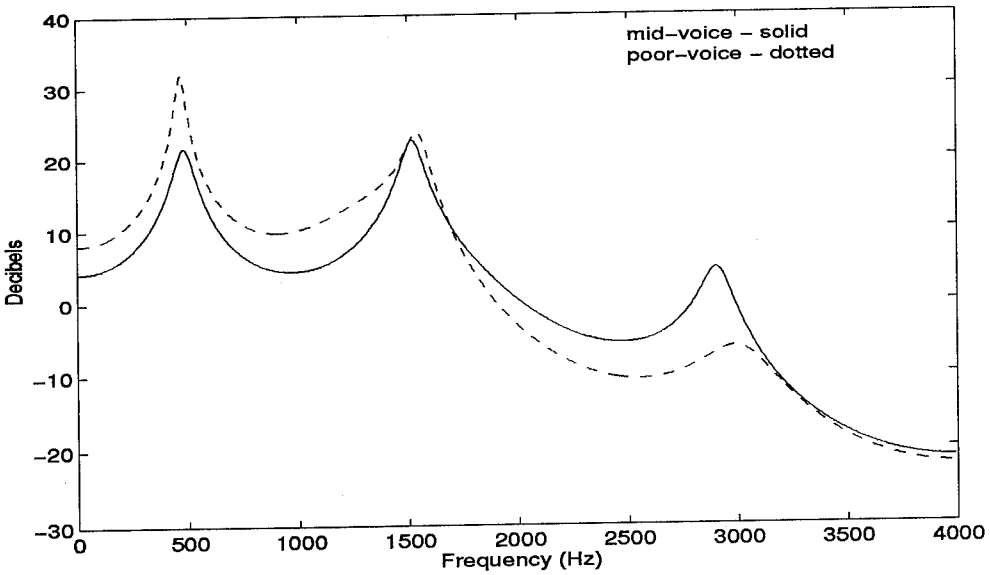


Figure 3.2: Spectral mismatch due to different channels.

dB for many recognition applications, the sensitivity in the performance accuracy is predominantly due to mismatched conditions between training and testing data due to convolutional distortions. Hence, it is of practical interest, to study the degradations caused by poor transmission channel conditions. In cases where the speech has been processed through a transmission channel or acquired via a microhone or a handset (for example in telephony applications), the transfer function corresponding to a speech segment can be simplified as,

$$S(z) = \frac{1}{A_{speech}(z)} \frac{N_{ch}(z)}{D_{ch}(z)}, \tag{3.2}$$

where $A_{speech}(z)$ represents an autoregressive fit solely due to speech in the frame while, $N_{ch}(z)$ and $D_{ch}(z)$ represent the zeros and poles of the overall channel distortion. The effect of noise has been left out in the formulation.

An AR fit to the spectra of a speech frame would yield a constrained all-pole representation of all zeros and poles of speech as well as of the channel. Thus, the parameters derived from the LP analysis not only represent the speech, but also perturbations caused by the transmission channel. Figure (3.2) illustrates an example of spectral mismatch caused to a frame of speech due to different transmission channels obtained from a simulator [88].

Since these perturbations vary with the channel, a mismatch occurs in evaluating the performance using these parameters. Moreover, such convolutional noise may affect each parameter differently. Parameters are sought wherein a simple transformation, or weighting attenuates the effect of convolutional distortions, thereby resulting in a more robust representation. In case of spectral modeling based on Fourier transform, the distortions affect the parameters (log spectral magnitudes) that are derived by sampling the spectrum of speech. It is possible to investigate the distortions caused by the transmission channels on model parameters in order to achieve channel normalization.

Channel invariant feature extraction algorithms in speech based systems have generally focussed on techniques that either,

- de-emphasize the effect of the channel on individual feature parameters, or,

- estimate and eliminate the convolutional distortion to obtain a robust feature representation.

Typically, parameters of speech are extracted using conventional signal modeling and then *modified* to yield an invariance to distortions. Compensation for convolutional distortions after feature extraction has a computational advantage since speech is processed in short segments. Channel compensation would be required only on those segments where there is necessary speech information while the rest of the signal can be eliminated from processing.

In practice, for spectral modeling of speech, the distortions caused by the communication channel or transducers are assumed to have a fixed frequency response (or time-invariant) and varying much more slowly than the speech signal itself.

A channel normalization approach would effectively deconvolve (inverse filter) the channel from speech. Channel normalization for recognition systems has frequently relied on a blind deconvolution approach based on homomorphic processing reviewed in Section (2.1.4).

For speaker recognition, features have generally been extracted using an LP derived cepstral transformation [1, 2, 31, 32]. For features based on Fourier transform modeling, the cepstral transformation has been proven to be superior for speech recognition applications. In contemporary recognition literature, feature extraction in the cepstral domain has widely been accepted as a standard. In fact, most investigative efforts in speaker as well as speech recognition systems have focused on features in the cepstral domain. The rest of the chapter reviews the significance of cepstral feature analysis and several related techniques that have been developed to minimize the mismatch caused by convolutional distortions.

3.2 Cepstral feature analysis

The cepstral parameterization of a short-time spectra of speech depends on the spectral modeling technique employed. In the case of LP modeling of the spectrum, if the all-pole filter $A(z)$ is stable (or minimum-phase), the cepstrum can be derived recursively. For a minimum phase all-pole filter that has all its roots inside the unit circle, $log(A(\frac{1}{z}))$ is analytic inside the unit circle and can be represented via a Laurent expansion ([13], pp. 230),

$$log(\frac{G}{A(z)}) = log(G) - log(A(z)) = log(G) - \sum_{k=1}^{\infty} c(k)z^{-k}. \tag{3.3}$$

Representing $A(z)$ in terms of its predictor coefficients,

$$A(z) = \sum_{k=0}^{P} a_k z^{-k} \quad a_0 = 1; a_P \neq 0, \tag{3.4}$$

and differentiating both sides by with respect to z^{-1} and then multiplying by z^{-1}, an expression can be derived that expresses, the cepstral coefficients in terms of the predictor coefficients. The predictor coefficients a_k and cepstral coefficients are related via a recursive relationship [1, 13],

$$\begin{aligned} c_1 &= -a_1; \\ c_n &= -a_n - \frac{1}{n}\sum_{k=1}^{n-1} k c_k a_{n-k} \quad 1 < n \leq P \\ c_n &= -\frac{1}{n}\sum_{k=1}^{n-1} k c_k a_{n-k} \qquad n > P. \end{aligned} \tag{3.5}$$

where P is the order of the filter and c_n is the n^{th} cepstral coefficient.

In case of Fourier Transform modeling of the spectra, the cepstral coefficients are calculated by first taking the *log* of the spectral magnitudes. An inverse Fourier transform ($IFFT \sim FFT^{-1}$) of the log spectrum yields the cepstral transformation as shown,

$$s(n) \Rightarrow FFT(s(n)) = S(\omega) \Rightarrow log(|S(\omega)|) \Rightarrow FFT^{-1} \Rightarrow c(n). \tag{3.6}$$

Conventionally the sequence $c(n)$ is truncated by windowing to a low order of 8–14 to form a feature vector, $\mathbf{c} = [c_1, c_2, \cdots, c_Q]$, of a speech frame for recognitionpurposes, where Q is the order of the cepstral sequence.

For speaker recognition purposes, LP-derived cepstral coefficients have a computational advantage due to the recursive relationship in equation [3.5], although either of the modeling techniques have shown comparable results in the cepstral domain [64].

Robustness techniques to convolutional distortions in the cepstral domain have relied on

1. Intraframe processing, and,
2. Interframe processing

of the cepstral parameters. The **Intraframe** approach involves modifying the cepstral features for an individual frame of speech to de-emphasize the coefficients that are sensitive to channel mismatch. The de-emphasis attenuates the mismatch in the cepstral parameters for a speech frame across varying distortions. The modification may either involve a weighting, selection or normalization. On the other hand, the **Interframe** approach involves investigating the time evolution of the parameters across an ensemble of speech frames or even the entire training or testing utterance. An estimate of the time-invariant distortion is often obtained by examining many speech frames representing the speech utterance. The estimate is then filtered or eliminated to generate robust parameters. Figure (3.3) illustrates the processing approaches. Both approaches are reviewed in the subsequent sections.

3.2.1 Conventional Intraframe processing

Cepstral coefficients used for recognition are generally weighted in order to minimize their sensitivity to channel differences, talker differences and ambient noise [11]. Conventional weighting schemes use a fixed cepstral window designed by studying the sensitivities of the cepstral coefficients. The first order cepstral coefficient, c_1, (c_0 correponds to the zeroth order coefficient) represents the tilt of the spectrum which is most drastically affected by differences in the channel characteristics. Lower order cepstral coefficients, in general, are more sensitive to channel differences. The higher order coefficients are sensitive to noise [11, 72, 73].

A simple cepstral weighting scheme applies an asymmetric triangular window, or a ramp weighting [72], given by,

$$\widetilde{c_m} = m c_m, \tag{3.7}$$

where m is the coefficient index. Ramp liftering (weighting in the cepstral domain) de-emphasizes the lower order cepstral coefficients which are more sensitive to channel variations. Linear weighting has the effect of a differentiator in the frequency domain or taking the first derivative of the log spectra. Distance measures based on the difference between the slopes of the log spectra have been shown to be robust to transmission channel variations [67]. A more complicated weighting scheme involves Bandpass liftering using a raised sine window,

$$\widetilde{c_m} = w_m c_m, \tag{3.8}$$

where,

$$w_m = 1 + \frac{Q}{2}\sin(\frac{\pi m}{Q}) \quad 1 \leq m \leq Q, \tag{3.9}$$

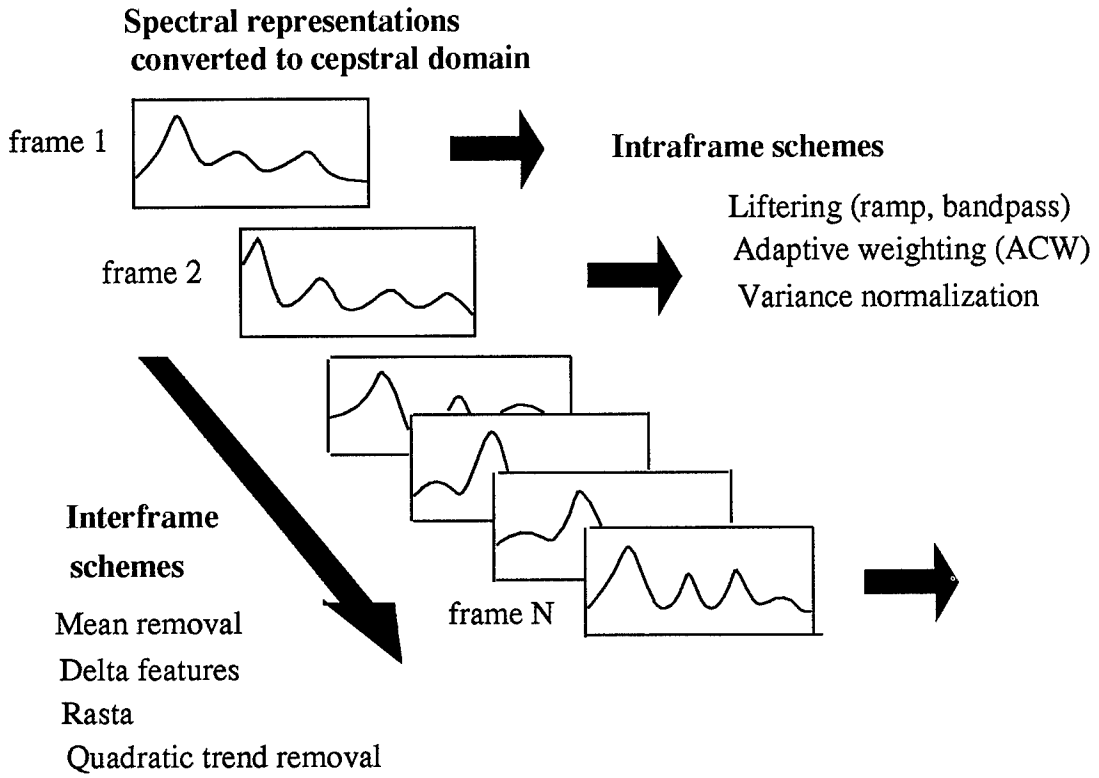


Figure 3.3: Processing of cepstral features.

where Q is the order of cepstral coefficients and the weighting, w_m, de-emphasizes the lower order cepstral coefficients to alleviate channel differences and the higher order cepstral coefficients to reduce the sensitivity to noise [73]. Figure (3.4) illustrates different lifters commonly employed for recognition purposes.

A more powerful intraframe approach to channel normalization was developed by Assaleh and Mammone [30, 45] wherein an adaptive cepstral weighting scheme was developed to minimize the effect of channel variations on the speech spectra. The approach derives a subtractive cepstral component for each frame of speech to attenuate channel differences. This approach will be unified in Chapter Four along with the Pole-filtering approach.

A cepstral weighting scheme based on the statistics of the variations in the cepstral parameters weights the cepstral coefficients by their individual variances over the test or training utterances. Cepstral coefficients have been observed to have variances that are inversely proportional to the square of the cepstral coefficient index, n, [73],

$$E\{|c_n|^2\} \propto \frac{1}{n^2}. \tag{3.10}$$

When cepstral vectors are used for comparing the training and testing utterance using a Euclidean distance metric, the result is likely to be dominated by terms that have large amplitude and variances. More often, the dynamic range and variance of the lower order cepstral coefficients is greater than the higher order cepstral coefficients. Weighting the cepstral coefficients by the variances, normalizes the contribution due to each dimension of the cepstral feature vector. Variance normalization also theoretically signifies a *Prewhitening Transformation* of the parameters, wherein the variances are assumed to correspond to the diagonal terms of a covariance matrix which are often used to decorrelate the parameters [15]. Since the cepstral features have generally been found to be uncorrelated (i.e, they correspond to a diagonal covariance matrix, with the off-diagonal terms being zero or a small value), a Euclidean distance metric

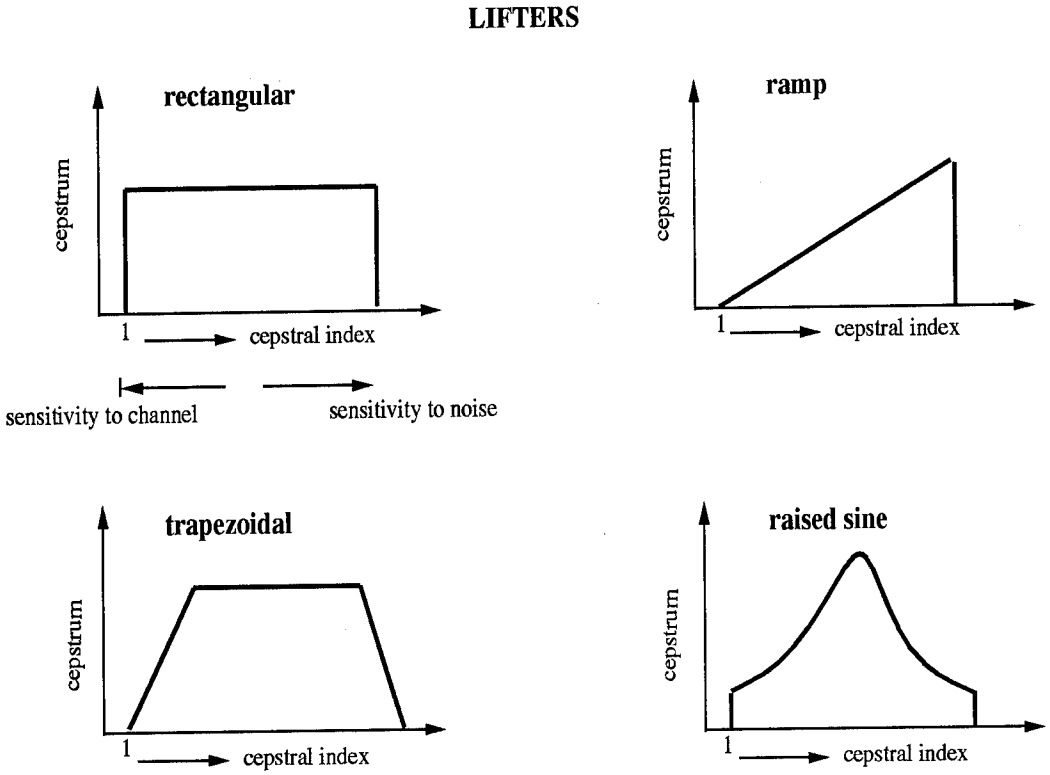


Figure 3.4: Liftering schemes on cepstral coefficients.

has often been a reasonable choice [7].

Intraframe techniques attempt to only minimize the sensitivity of the cepstral features to convolutional distortions, but do not necessarily eliminate them. Interframe approaches on the other hand explicitly attempt to filter or eliminate the contribution due to channel from each frame of speech. Such estimation of the components due to the channel is discussed in the following section.

3.2.2 Conventional Interframe processing

Interframe processing exploits the temporal variability of a sequence of cepstral vectors. In practice, the sequence may correspond to a training or testing utterance recorded via a microphone and transmitted over a comunication channel. Interframe processing is invariably based on the assumption that the convolutional noise is slow-varying or time-invariant over the interval of analysis. The transmission channel and microphone are assumed to have a fixed frequency response over the duration of the training or testing utterance. Based on this assumption, cepstral domain processing renders the time-invariant convolutional distortion as a *constant (dc) bias* (refer to section 2.1.4).

An estimate of this *bias* can be estimated by averaging in the cepstral domain and then subtracting this average component from the cepstral vector of each frame. All interframe techniques in the cepstral domain implicitly utilize this estimate as the premise for *Channel normalization.*

Cepstral Mean Normalization (CMN)

CMN [1], proposes that the effect of any fixed frequency response distortion introduced by the recording apparatus or the transmission channel can be eliminated from the cepstral sequence of a speech utterance by subtracting its long-term mean. The process is also referred to as Cepstral Mean Subtraction (CMS).

If a speech utterance S, consists of M overlapping short-time segments, then long-term mean of the cepstral coefficients is given by,

$$\mathbf{c}_S = \left(\langle \tfrac{1}{M} \textstyle\sum_{m=1}^{M} c_1(m) \rangle \quad \langle \tfrac{1}{M} \textstyle\sum_{m=1}^{M} c_2(m) \rangle \quad \cdots \quad \langle \tfrac{1}{M} \textstyle\sum_{m=1}^{M} c_Q(m) \rangle \right), \tag{3.11}$$

where m is the frame index, and Q is the order of dimensionality of the cepstral vectors. The cepstral mean normalization for the m^{th} frame of speech is given by,

$$\mathbf{c}_m - \mathbf{c}_S, \tag{3.12}$$

where $\mathbf{c}_m$ is the cepstral sequence for the m^{th} frame.

Since the long-term cepstral average represents the cepstral estimate of the convolutional distortion, it is also frequently called the *Channel Cepstrum*. Although simple, CMN has been proven to be the single-most powerful method for normalizing channel differences and minimizing the mismatch between training and testing utterances [80, 64, 31, 32]. CMN, however, has major shortcomings when the speech data available for training and testing is limited (also discussed in Section 2.1.4, equation [2.15]). This is due to the short utterance duration, cepstrum corresponding to the underlying speech tends to have an invariant component. This invariant component relates to the gross spectral distribution of the speech utterance. This aspect will be analyzed in greater detail in Chapter four wherein the spectral distribution of the cepstral mean will be investigated.

To verify that, for short utterances, the cepstrum corresponding to clean speech has an invariant component or that it is not zero-mean, an experiment was performed on clean speech from the KING database [89]. The cepstral mean for various utterance durations [1] is plotted in Figure (3.5) for a male speaker from the KING database. The figure represents the cepstral mean coefficients (12^{th} order) of a clean speech utterance [2] (obtained from the wideband speech portion) and the same utterance degraded by a real telephone channel (obtained from the narrowband portion). It can be observed that the cepstral features are not zero mean but constitute a bias that corresponds to clean speech representing a gross spectral distribution for a speaker. One can also observe that although the cepstral coefficients for clean speech may tend to zero over a very large time durations [73], this property is not apparent for short utterances.[3]. A discussion of the zero-mean property of cepstral coefficients is also discussed in Appendix B. An alternate method for proving that cepstral mean of short duration clean utterances is not zero-mean was also tried wherein speaker recognition experiments were conducted of cepstral features of clean speech and repeated on the same cepstral features after CMN. The speaker recognition accuracy was found to degrade by twenty percent when the cepstral mean for clean speech utterances was eliminated. When CMN is carried out in the presence of a convolutional distortion, this underlying invariant component due to clean speech is also eliminated from each frame. Clearly, an inaccurate estimate of the channel cepstrum, biased by an invariant cepstral component due to speech is eliminated. In order to maximize the discriminability among the speakers, the information that corresponds to the cepstrum of the gross spectral distribution of speech must be retained.

Hence, although CMN performs a reasonably good job in normalizes the channel mismatch between training and testing data, it tends to attenuate speech information when subtracted from the cepstral vector corresponding to a speech frame.

For short utterances, the presence of invariant speech information in the cepstral mean causes an elimination of useful spectral information from the cepstrum of an individual speech frame.

[1] Representative of durations for training and testing in many practical applications.

[2] The notion of a clean speech utterance in this thesis implies speech collected using a high quality microphone and not acquired over a telephone channel.

[3] Short utterances are considered to be of the order of one to ten seconds in this thesis. All simulations and experiments have focussed only on durations of this order

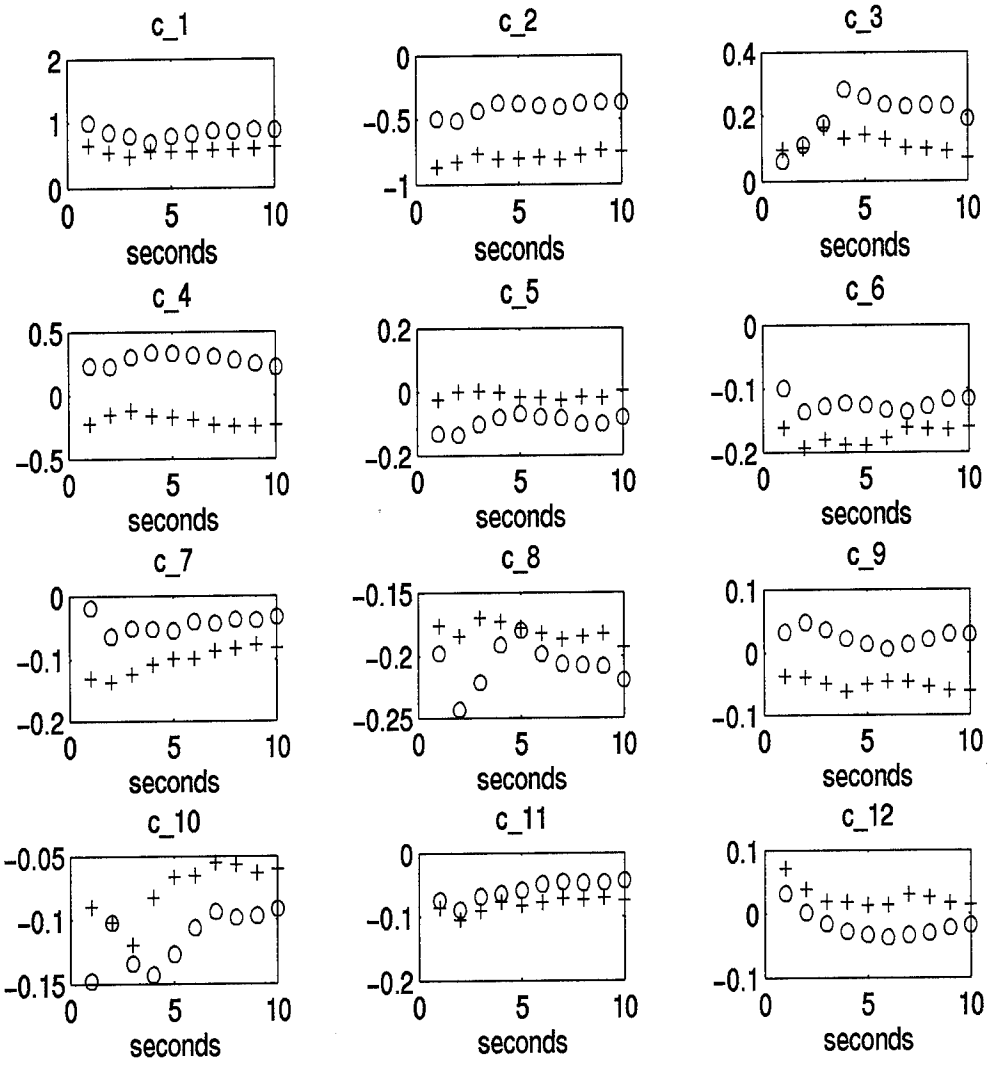


Figure 3.5: Experiment to prove cepstral features for clean speech are not zero-mean for short utterances, 'o' represents clean speech, whereas, '+' represents telephone channel speech.

The degradation in the performance accuracy due to CMN has been addressed in the past when the training and testing data were subjected to same convolutional distortion [4, 28] for speaker recognition. Recently, this issue has also been addressed critically for speech recognition by Neumeyer et.al. [80]. This effect will be discussed later in detail in Section (4.3.1), wherein an improved estimate of the mean will be developed using pole-filtering.

It should be noted that subtracting the long-term mean in the cepstral domain is equivalent to dividing by the geometric mean in the spectral domain. Hence the effect of subtraction in the cepstral domain, effectively filters the spectral information in the frequency domain as a blind deconvolution process (refer to Appendix A).

Dynamic cepstral features

The dynamic aspects of features are often represented by a temporally differentiated feature set. One approach was suggested by Furui [32], wherein the temporal sequence of the cepstral features was approximated by a polynomial approximation. This approximation has the effect of bandpass filtering the time trajectories of the cepstra. The filtered coefficients are called delta-cepstral coefficients and are given by,

$$\triangle c_n(m) = \sum_{k=-K}^{k=K} c_n(m-k)\delta(k), \tag{3.13}$$

where m is the cepstral frame index, and n is the feature dimension index. $\delta(k)$ represents the impulse response corresponding to a $2K+1$ tap bandpass filter which approximates the derivative of $c_n(m)$. The filter taps are given by:

$$\delta(k) = \frac{k}{\sum_{k=-K}^{K} k^2}, \tag{3.14}$$

where K takes typical values of 2 or 3 [11].

The effect of such bandpass filter is also to eliminate the dc-component of the log spectra. Hence the delta-cepstral feature implicitly performs a CMN. It has been shown that such dynamic features improve the performance only when they are concatenated with the static cepstral features of each frame of speech. The static cepstral features, however, need to be normalized via CMN before the dynamic features are appended. The incorporation of dynamic features has been shown to improve the performance of most speech recognition systems. This is due to the fact that the derivative emphasizes the fast varying spectral transitions in speech [32, 35].

RASTA processing approach

A related interframe processing technique which has recently gained more prominence is known as RASTA (RelAtive SpecTrA) [68]. In a manner similar to the delta-cepstrum, RASTA has the effect of bandpass filtering the time trajectories of the cepstral coefficients. This filtering effectively eliminates the temporal average of the cepstral sequence. The high pass portion of the bandpass filter is expected to alleviate the effect of the convolutional distortions, whereas the low pass portion smoothes some of the fast frame-to-frame spectral changes in the short-term spectral estimate attributed mainly due to the analysis artifacts [79].

The RASTA LP cepstrum $\triangle_R c_n(m)$ is given by,

$$\triangle_R c_n(m) = \sum_{k=-K}^{K} c_n(m-k)\delta(k) + \alpha \triangle_R c_n(m-1), \tag{3.15}$$

where α corresponds to the tap of a first order autoregressive filter.

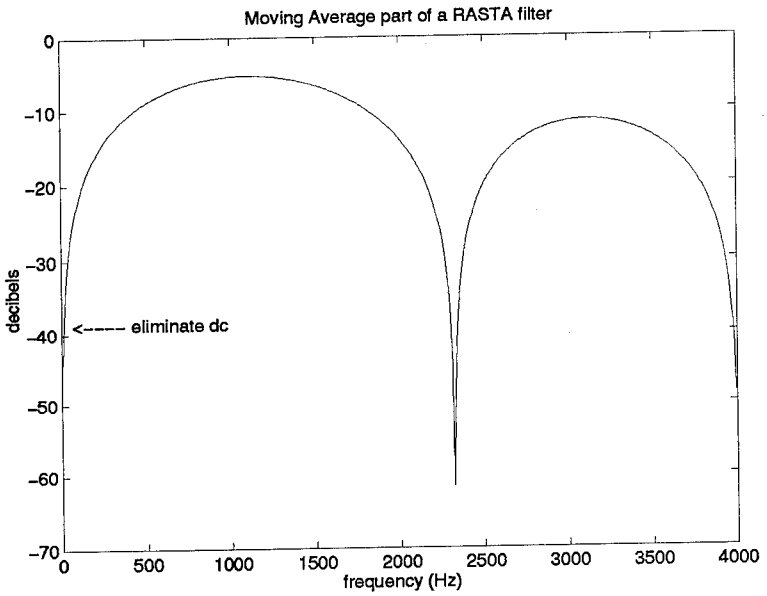


Figure 3.6: A typical RASTA filter.

The moving average (MA) part of a typical RASTA filter ($K = 2$) that helps minimize the convolutional distortion by eliminating the dc component in the log spectral domain is,

$$MA(z) = z^4(0.2 + 0.1z^{-1} - 0.1z^{-3} - 0.2z^{-4}). \tag{3.16}$$

The frequency response of the filter is shown in Figure (3.6).

Since RASTA processing applies the filter locally, besides making the long-term average of the log-spectrum identically zero, it also combines a weighted average of all the past log-spectra with the current log-spectrum thus allowing an exponentially decaying mean of all past analysis frames with the use of an autoregressive portion. Such exponentially decaying running average can follow the slow changes in the communication environment. For speaker recognition however, it has been found that with local frame differencing carried out by RASTA (similar to computing delta-features), an unreliable channel estimate is eliminated from the cepstral vector [64]. It is conjectured that RASTA processing implicitly suffers from the same drawback as ordinary CMN for short duration utterances.

A RASTA processing on cepstral trajectories is illustrated in figure (3.7).

Combined Interframe and Intraframe approaches

Channel normalization techniques proposed in the past have evaluated interframe and intraframe approaches independently. While intraframe approaches have focused on schemes to enhance spectral information relating to speech in the presence of distortions, interframe techniques emphasize the elimination of degradations due to the channel for computation of robust cepstral features. Several hybrid approaches have been proposed which combine intraframe techniques typically followed by cepstral mean removal to eliminate the time-invariant distortion. In this report a new hybrid approach is proposed wherein an estimate of the channel distortion is deconvolved from the speech signal in a first pass followed by intraframe processing to further minimize the channel mismatch. The combined processing, although involves a computationally expensive two-pass approach, is shown to perform better than individual interframe and intraframe approaches for channel normalization. This approach shall be outlined in detail in Chapter four.

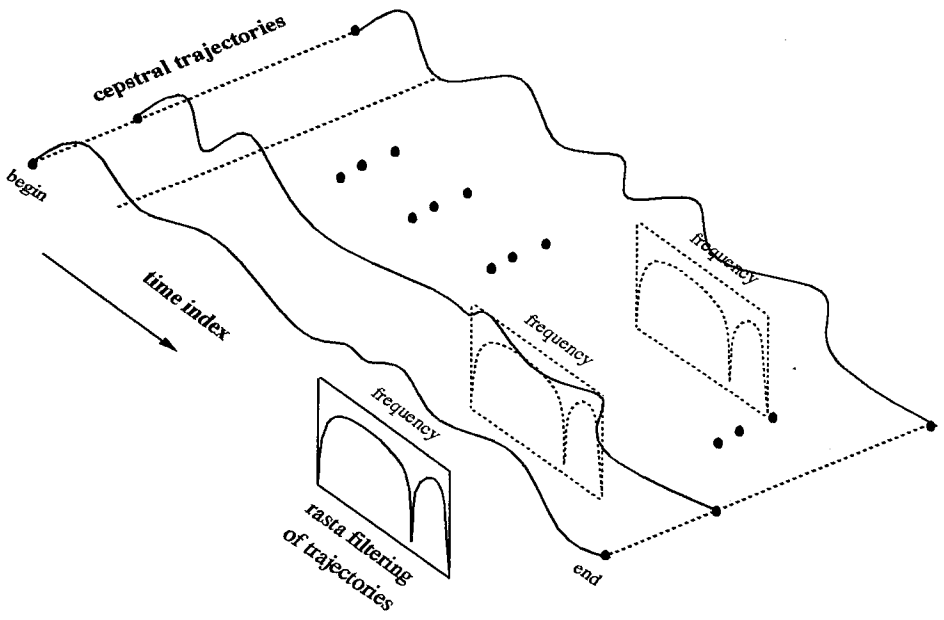


Figure 3.7: RASTA processing of cepstral trajectories.

The subsequent chapter introduces the pole-filtering approach and establishes the basis for improved channel normalization over conventional techniques. The need for an improved cepstral mean estimate also justifies improving all conventional approaches that implicitly perform cepstral bias elimination with a refined cepstral bias.

3.3 Probabilistic Channel normalization

Several methods have been developed in the past to statistically model and estimate the channel in a probabilistic framework [28, 74, 80, 77, 76]. The channel is modeled as a Gaussian random vector and the information is incorporated in a maximum likelihood framework. Such probabilistic modeling is easily integrated into a Hidden Markov Model or a Gaussian Mixture Modeling framework for classification. The mismatch, due to channel differences between training and testing, may be minimized in the cepstral domain by [74],

- estimating the bias due to the channel component, or,
- adapting the testing conditions to that of the trained models using a stochastic mapping approach.

In a probabilistic framework, the cepstral bias is eliminated by maximizing the likelihood of a speech model wherein the cepstral bias is considered an unknown parameter. Another method estimates the bias by evaluating a maximum-likelihood channel estimate given a collection of observations and model parameters [28, 51, 80]. In either case, the maximum-likelihood estimate of the speech model or the channel is obtained using an iterative Estimation-Maximization (EM) [95] algorithm.

The stochastic matching method [77] maps the distorted features $\mathbf{Y}$ corresponding to the testing conditions to an estimate of the features $\mathbf{X}$ that correspond to the training conditions such that $\mathbf{X} = F_\nu(\mathbf{Y})$.

The matching between the training and testing is carried out by estimating the unknown parameter ν iteratively using an EM algorithm, to maximize the likelihood of the observed features $\mathbf{Y}$, given the models corresponding to features $\mathbf{X}$.

Probabilistic channel modeling and adaptation has been prevalent in speech recognition applications that frequently employ a Hidden Markov Modeling framework. However, these techniques rely on a notion of a cepstral bias that is different from the cepstral mean bias employed in ordinary CMN. The cepstral bias is defined when a mismatch exists between a training model and the testing data. Thus, the bias may be considered as zero-mean for training and testing on clean speech and non-zero quantity to be estimated using Maximum-Likelihood estimation in case of environmental degradations. Probabilistic modeling techniques have been shown to perform comparably or sometimes better than cepstral mean subtraction [74, 80]. Parametric modeling for speech and speaker recognition generally requires substantial training data to model the underlying statistics of the speech or speaker properly. This report focusses on acquiring improved channel estimates for short training and testing durations and hence, only non-parametric modeling such as VQ shall be emphasized.

Chapter 4

Feature extraction based on Pole-filtering

A new methodology for extracting robust features for speaker recognition is introduced in this chapter. The methodology called **Pole filtering**, is shown to yield features that are robust to channel differences in performing a speaker recognition task. The Pole-filtering approach is used to perform channel normalization using intelligent filtering of the eigenmodes of speech corresponding to the pole parameters of the all-pole LP filter. Channel normalization is achieved via a transformation to the cepstral domain, wherein an improved estimate of the channel cepstrum is obtained by studying the effect of channel variations on eigenmodes of speech.

Intrinsic to this philosophy is an explanation why the LP derived cepstral transformation of speech is a powerful feature set for recognition systems. The approach is outlined by introducing the concept of natural modes of vocal tract in section (4.1), followed by the relationship between the cepstrum and the natural modes in section (4.2). The effect of channel variations on modes of vocal tract and coresponding cepstral transformation is then investigated in section (4.3). Based on this investigation, an Interframe pole filtering approach is developed. A previously developed Intraframe processing approach [30, 31] is unified under the pole-filtering methodology. A general method of achieving improved channel normalization by utilizing channel estimates from the interframe approach followed by conventional intraframe processing is introduced at the end of the chapter.

4.1 Modeling based on eigenmodes of speech

A segment of speech can be modeled by a difference equation. The natural modes of the resulting linear time–invariant system can be obtained in the following way. The difference equation for stationary frame of speech can be reformulated as,

$$\sum_{k=0}^{P} a_k s(n-k) = 0 \quad a_0 = 1; \tag{4.1}$$

where $s(n)$ is the speech sample and a_k are the predictor coefficients.

From linear system theory, the homogeneous solution of this difference equation is given by [19],

$$s_h(n) = \sum_{k=1}^{P} b_k z_k^n, \tag{4.2}$$

where s_h is the homogeneous solution and, $z_k, k = 1, 2, \cdots, P$, are the distinct roots of the difference equation. The constants b_k are evaluated from the initial conditions for the linear system. Hence the

general solution is a weighted power sum of the natural modes of the system that correspond to the roots of the difference equation.

For a speech frame,

$$S(z) = \frac{1}{A(z)} = \frac{1}{1 + \sum_{k=1}^{P} a_k z^{-k}}, \tag{4.3}$$

where,

$$A(z) = \sum_{k=0}^{P} a_k z^{-k} \quad a_0 = 1; a_p \neq 0. \tag{4.4}$$

The transfer function can be viewed as an all-pole filter of order P with roots $z_k, k = 1, 2, \cdots, P$ of $A(z)$. Thus the roots z_k, correspond to the modes of the linear system of speech. These roots form the dominant modes in the speech segment. Each root z_k has associated with it a bandwidth, B_k, and a center frequency, ω_k, given by the relation,

$$z_k = e^{-B_k + j\omega_k}, \tag{4.5}$$

where,

$$\omega_k = \frac{1}{2\pi} \arctan \frac{\Im(z_k)}{\Re(z_k)}, \tag{4.6}$$

and

$$B_k = -\frac{1}{\pi} ln(|z_k|), \tag{4.7}$$

in units of π radians.

It is possible to interpret equation [4.3] in various forms such as,

$$S(z) = \frac{1}{A(z)} = \frac{1}{1 + \sum_{k=1}^{P} a_k z^{-k}} = \frac{1}{\prod_{k=1}^{P} (1 - z_k z^{-1})} = \sum_{k=1}^{P} \frac{r_k}{(1 - z_k z^{-1})}, \tag{4.8}$$

where z_k are the poles of the all-pole filter. Figure (4.1) shows the interpretation of poles on the unit circle (Z–domain).

Thus, the linear system modeling the speech segment can be considered as a cascade of P (being the order) first order filter sections having transfer function $\frac{1}{1 - z_k z^{-1}}$. A partial fraction expansion leads to a parallel form representation weighted by their corresponding residues, r_k, which are evaluated using,

$$\lim_{z \to z_k} \langle \frac{(1 - z_k z^{-1})}{A(z)} \rangle. \tag{4.9}$$

The roots of the all-pole filter $A(z)$ either occur in complex conjugate pairs or are real roots. An all-pole filter having P poles, may have q pairs of complex poles and the remaining $(P - 2q)$ real poles. The impulse response of a complex conjugate pole pair corresponds to a damped sinusoid at an angular frequency $\omega_k = \frac{1}{2\pi} \arctan \frac{\Im(z_k)}{\Re(z_k)}$ and a damping factor corresponding to $|z_k|$. The resulting impulse response is given by,

$$\frac{1}{(1 - z_k z^{-1})(1 - z_k^{\star} z^{-1})} \xrightarrow{Z^{-1}} |z_k|^n \cos(\omega_k n). \tag{4.10}$$

Each complex conjugate pole pair represents a component in the spectral domain (referred to as a **spectral component**) corresponding to a frequency ω_k, and bandwidth B_k, where $k = 1, 2, \cdots, q$. The real poles correspond to roots on the real-axis of the unit circle at frequencies of $\omega_k = \langle 0, \pi \rangle$. The poles of

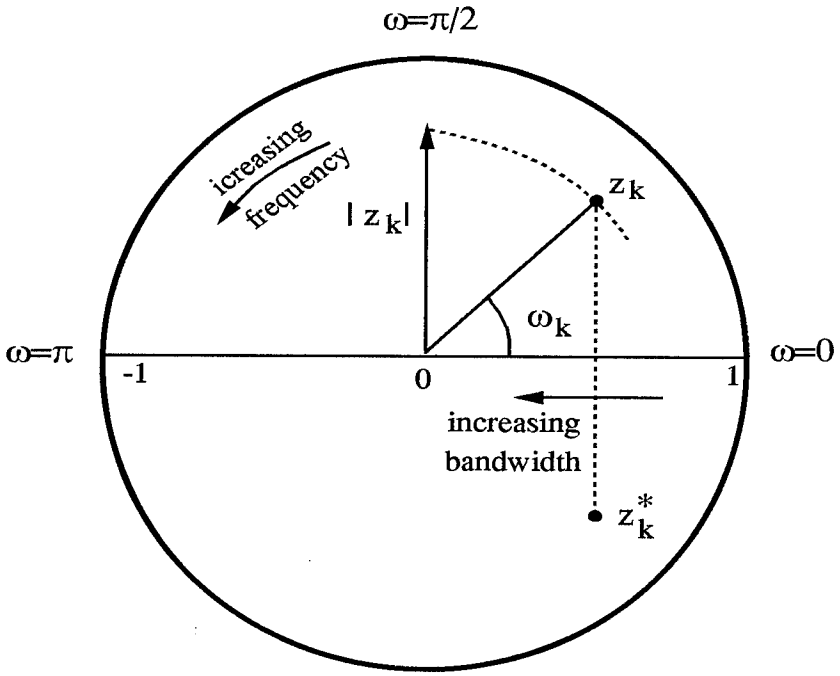


Figure 4.1: Pole domain interpretation.

the all-pole filter closer to the unit circle correspond to narrow-band components with smaller bandwidths and larger decay factors $|z_k|$ (i.e, $|z_k| \rightarrow 1$), while the poles closer to the origin of the unit circle correspond to large bandwidth components and have a small decay factor ($|z_k| \rightarrow 0$). The real poles correspond to the spectral tilt. Figure (4.2) illustrates components in the $\mathcal{Z}$-domain, frequency domain and the time domain.

For relatively clean speech, some of the dominant modes often represent the formant modes of speech that correspond to components with narrower bandwidths. Broad-band components corresponding to real poles attempt to model the spectral tilt, sub-glottal effects etc. In the case of modeling practical speech signals, the transmission channel and ambient noise have an adverse influence on the natural modes of speech. The all-pole model then represents modes which are perturbed by the transmission channel and noise. In addition to this, under low SNRs, spurious poles tend to appear in the all-pole fit to the speech spectra. A study of how channel degradations affect the modes and the components representing a frame of speech forms the basis of the pole-filtering methodology developed later in this chapter. It should be noted that the effect of noise on the components is not investigated in this report.

4.2 Relationship betweem cepstrum and modes of speech

The LP derived cepstrum from equation [2.11] can be seen to be as being the inverse transform of the natural logarithm of the short-time LP transfer function, $S(z)$. It is the impulse response of $ln(S(z))$ which is given by,

$$ln(S(z)) = \sum_{n=1}^{n=\infty} c_n z^{-n}, \tag{4.11}$$

where c_n is the n^{th} cepstral coefficient.

The predictor coefficients a_k and cepstral coefficients are related via a recursive relationship given by equation [3.5]. Alternately, the cepstral coefficients can also be obtained by relating them to the poles

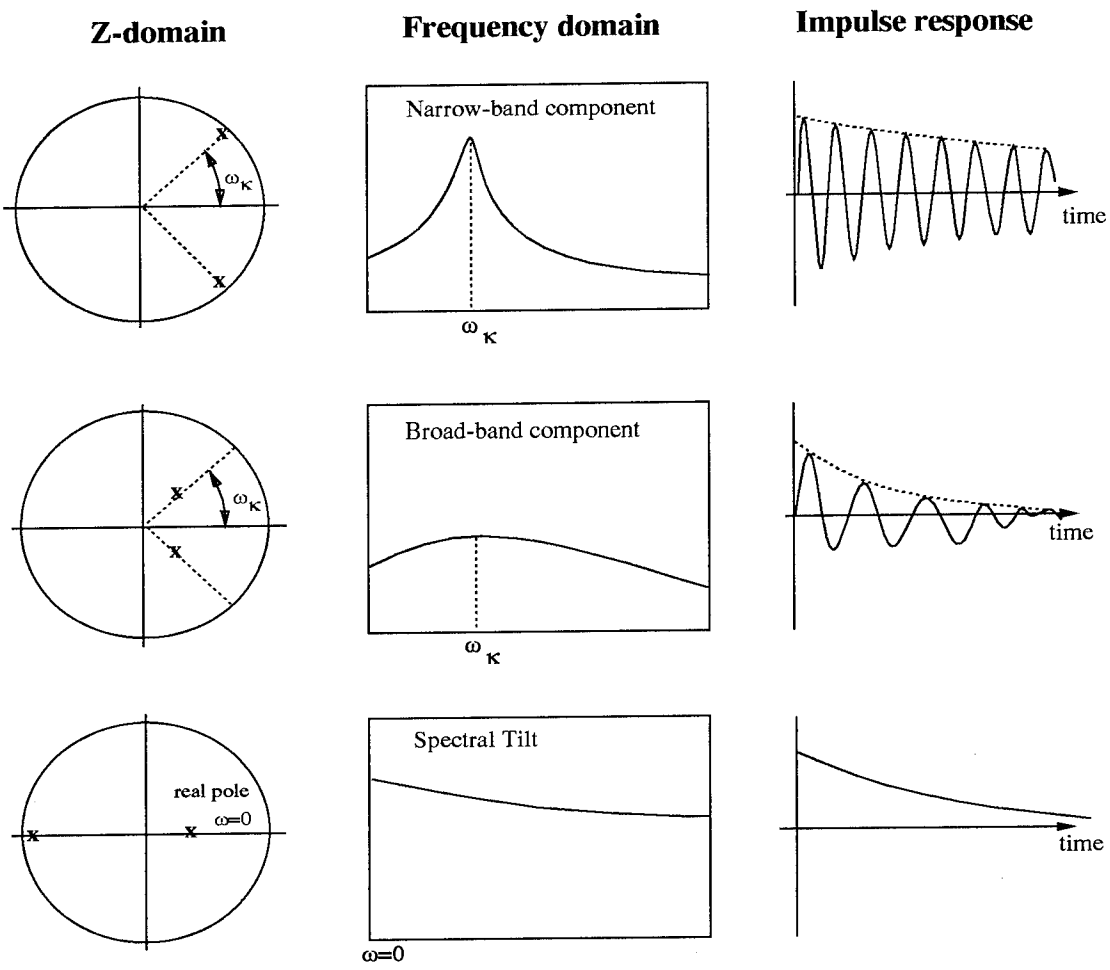


Figure 4.2: Meaning of components in Z-domain, Frequency and Time domains.

of $S(z)$ and hence to the center frequencies and bandwidths. Substitution of equation [4.8] into equation [4.11] yields,

$$\sum_{k=1}^{P} \ln(1 - z_k z^{-1}) = -\sum_{n=1}^{\infty} c_n z^{-n}. \tag{4.12}$$

The factor $\ln(1 - z_k z^{-1})$ can be expanded [106] as,

$$\ln(1 - z_k z^{-1}) = -\sum_{n=1}^{\infty} \frac{1}{n} z_k^n z^{-n}. \tag{4.13}$$

By combining equations [4.12] and [4.13], the cepstral coefficients, c_n, can be expressed in terms of the roots of the LP polynomial or poles of the all-pole filter as,

$$c_n = \frac{1}{n} \sum_{k=1}^{P} z_k^n. \tag{4.14}$$

Thus c_n can be interpreted as the power sum of the LP polynomial roots normalized by the cepstral index [87].

Comparing equation [4.14] to the homogeneous solution of the difference equation of speech in equation [4.2], one can see that the liftered cepstral sequence is a special case of the homogeneous solution, with the coefficients $b_k, k = 1, 2, \cdots, P$ set to unity.

The relationship of the cepstrum to the natural modes of speech can also be expressed as [31],

$$\begin{aligned} c_n &= \frac{1}{n} \sum_{k=1}^{P} e^{-n(B_k + j\omega_k)} \\ &= \frac{1}{n} \sum_{k=1}^{\frac{P}{2}} e^{-nB_k} cos(n\omega_k). \end{aligned} \tag{4.15}$$

Thus, the n^{th} cepstral coefficient can be interpreted as a nonlinear transformation of the resonance center frequencies and bandwidths.

More importantly the LP derived cepstral coefficients also form a special case of the homogeneous solution of speech. These relationships give us a physical insight into why the cepstrum offers a powerful feature set. On the other hand, with the root power sum formulation of the cepstrum one can study the effect of each spectral component and hence each pole on the overall cepstrum of a speech frame. A weighting scheme that de-emphasizes the contribution of spectral components that are sensitive to environmental degradations to form a robust cepstral feature set is implied. A study of how each spectral component affects the cepstral transformation can be used to design robust feature extraction schemes.

4.3 Pole-filtering methodology

The process of *filtering, selecting* or *weighting* the poles of a speech frame and their respective spectral components or their parameter transformations will be called **Pole Filtering** in this context. The basis for the pole-filtering approaches for robust channel normalization is established by studying

1. the effect of poles of speech on *apriori* known channel distortions, and,
2. the effect of channel distortions on the poles of speech.

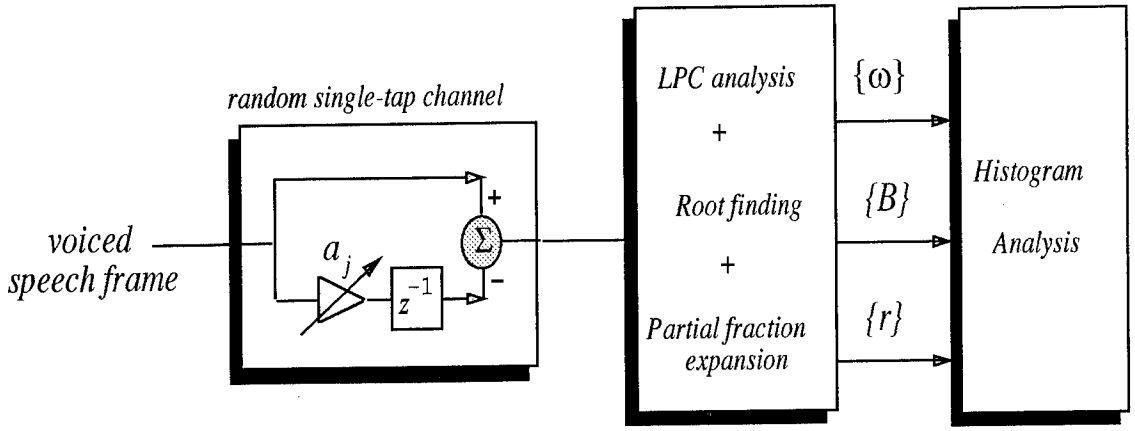


Figure 4.3: Block diagram of the channel variation experiment (adapted from Assaleh and Mammone [31]).

The study of the effect of poles of speech on the cepstral estimate of the convolutional distortion degrading a speech utterance forms the premise of the Interframe Pole-filtering approach. An investigation of the effect of channel variations on the poles forms a basis for the Intraframe pole filtering approach.

The study of channel variations on the different natural modes of speech and the poles of an LP filter modeling the transfer function for a known frame of speech was carried out by Assaleh and Mammone [31]. The spectral components, z_k, parameterized by their respective center frequency ω_k, a bandwidth B_k, and residues r_k, were studied under perturbations caused by channel distortions. Each of these parameters were found to be sensitive to channel variations. The sensitivities of these parameters were investigated experimentally. The highlights of the findings will be reviewed here to aid the development of the pole-filtering methodology.

The experiment illustrated in Figure (4.3), was carried out as follows,

- A voiced frame of speech was processed through a random single-tap channel given by

$$\Theta_j(z) = 1 - a_j z^{-1}, \tag{4.16}$$

  where a_j is a sequence of uniformly distributed random numbers between 0.0 and 1.0.

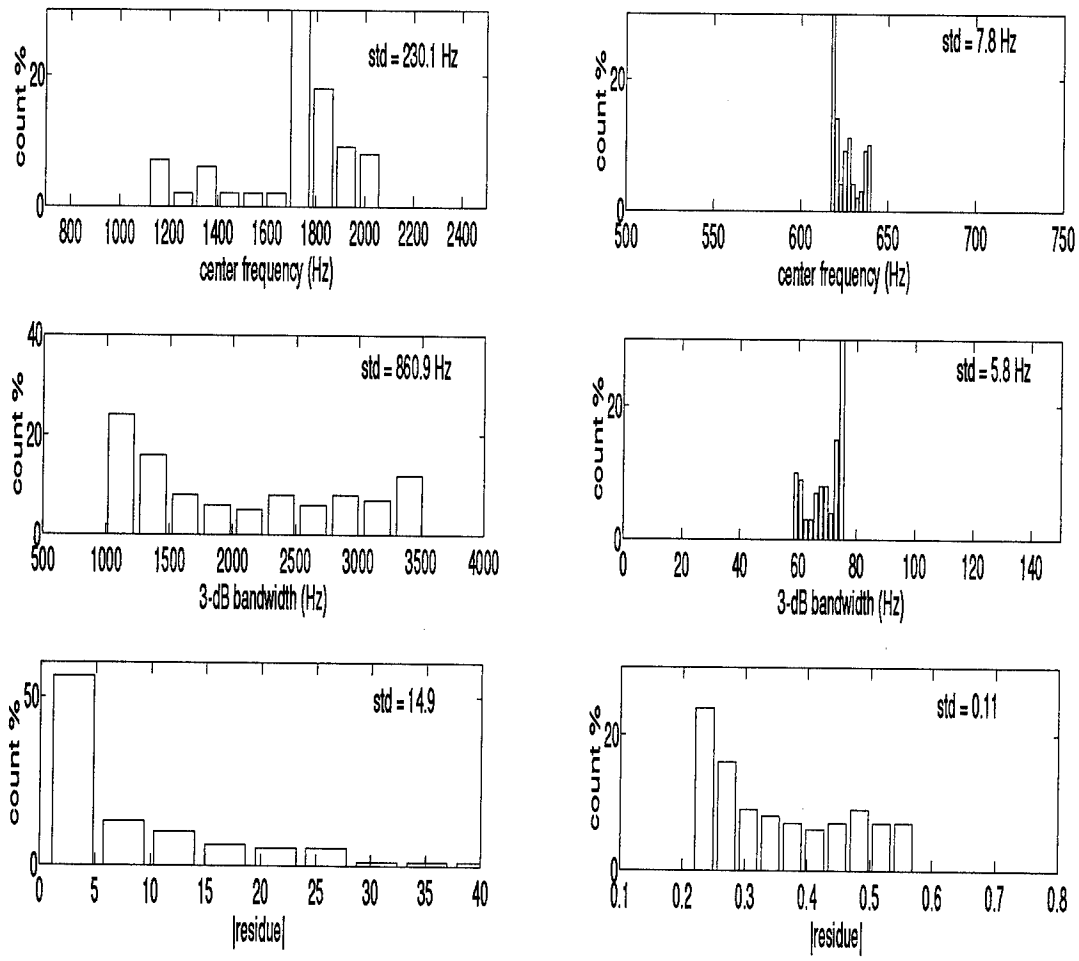
- The sequences of the parameters $(\omega_i,\ B_i,\ r_i)$ of all components were computed for each j in the random sequence a_j.

- Two sequences of $(\omega_i,\ B_i,\ r_i)$ were selected to represent a narrow-bandwidth component and a broad-bandwidth component.

- The sensitivity of the parameters of the selected narrow-bandwidth and broad-bandwidth components were evaluated by histogram analysis.

By examining the resulting histograms of the parameters of the broad-bandwidth component shown in Figure (4.4), it was concluded that the three parameters,$(\omega_i,\ B_i,\ r_i)$, associated with such components possess large variances with respect to channel variations, hence are very sensitive to channel differences.

Narrow-bandwidth components were shown to preserve their center frequencies and bandwidths since their histograms exhibited small variances. Narrow-band components, hence represent the stable modes of speech, which are more robust to channel variations. The residues however, demonstrated large variance even for narrow-band components. Observations made from this experiment were used previously to



(a) (b)

Figure 4.4: Histograms of the parameters of (a) a broad-bandwidth component and (b) a narrow-bandwidth component (adapted from Assaleh and Mammone [31]).

develop an intraframe adaptive component weigthing method which will be reviewed and unified under the pole–filtering methodology later in the chapter.

Alternately, observations made in this experiment motivated the study of a variant of the experiment. The variant consists of studying the effect of narrow-band and broad-band components of speech on the cepstral estimate of an *apriori* known channel distortion. An interframe pole-filtering of the channel cepstrum estimate has been developed for improved channel normalization. The approach is explained in the following section.

4.3.1 Interframe Pole-filtering approach

The Interframe pole-filtering approach exploits the root power sum relation between the cepstrum and the poles of the all-pole LP filter stated in equation [4.14]. The long-term cepstral mean [Sections (2.1.4) and (3.2.2)] for an utterance is analyzed in order to develop the approach.

Analysis of CMN

The long-term cepstral mean for an utterance, S, was shown to provide an estimate of the channel in Section (2.1.4). The channel cepstrum, $\mathbf{c}_S$, corresponds to a sum of the cepstral vectors for individual frames of speech of the utterance S. If $\mathbf{s}_S$ corresponds to the cepstral component due to clean speech prior to being degraded by the convolutional distortion, and $\mathbf{h}$, the actual channel cepstrum estimate, then,

$$\mathbf{c}_S = \mathbf{s}_S + \mathbf{h}. \tag{4.17}$$

The channel cepstrum for the utterance S using a root power sum interpretation is given by,

$$\mathbf{c}_S = \left(\frac{1}{M}\sum_m \langle \sum_{k=1}^{P} z_{k,m}^{1} \rangle \quad \frac{1}{M}\sum_m \langle \sum_{k=1}^{P} z_{k,m}^{2} \rangle \quad \cdots \quad \frac{1}{M}\sum_m \langle \sum_{k=1}^{P} z_{k,m}^{Q} \rangle \right), \tag{4.18}$$

where m is the frame index for the M frames of the utterance S, the cepstral indices vary from $1, \cdots, Q$ for each of the Q dimensions of the cepstral vector $\mathbf{c}_m$ and P is the order of the LP filter which corresponds to the number of roots, z_k, in the root power sum formulation. Frequently the order of the LP fit and that of the cepstral sequence are kept the same, i.e $P = Q$.

Alternately, the channel cepstrum can be represented by,

$$\mathbf{c}_S = \sum_{m=1}^{M} \mathbf{s}_m + \mathbf{h}, \tag{4.19}$$

where $\mathbf{s}_S = \sum_{m=1}^{M} \mathbf{s}_m$ corresponds to the cepstral mean component solely due to underlying clean speech. As mentioned in Section (2.1.4), the component due to clean speech should be zero-mean in order for the channel cepstrum estimate $\mathbf{c}_S$ to correspond to cepstral estimate, $\mathbf{h}$, of the actual underlying convolutional distortion.

It was empirically shown in section (3.2.2) that the mean cepstrum component due to clean speech is never zero for short utterances. This may often be the case for training and for testing. That is, the channel cepstrum consists of an invariant component due to speech which is eliminated when ordinary CMS is performed. Since the channel cepstrum now contains a gross spectral distribution due to channel as well as speech, the elimination of a distorted estimate of the channel cepstrum from cepstrum of each speech frame corresponds to effectively deconvolving an unreliable estimate of the channel. In other words, spectral components that implicitly corresponds to the gross spectral distribution of clean speech is also deconvolved from the speech utterance.

An estimate of the channel cepstrum biased by the invariant component due to speech impairs proper channel compensation in the cepstral domain. A considerable loss in recognition accuracy is observed in speaker recognition systems when the system is trained and tested on similar or different channels[4, 28].

The effect of ordinary cepstral mean removal can be illustrated by introducing the concept of a cepstral Channel Compensation Filter (CCF).

Channel Compensation Filter (CCF)

The spectral distribution of the channel cepstrum, $\mathbf{c}_S$, can be analyzed by observing the frequency response of a Channel Compensation filter. One can observe from the expression in equation [3.5], that the predictor coefficients can be transformed into cepstral coefficients recursively. Similarly, the prediction coefficients can be recursively obtained from the cepstral coefficients. The prediction coefficients, derived from the channel cepstrum, correspond to a Channel Compensation filter (CCF) which represents an all-pole approximation of the convolutional distortion.

Since the long-term cepstral mean of the utterance is obtained from a summation of causal cepstral sequences [1] that correspond to each frame of the speech utterance, the channel cepstrum can be assumed to be causal. A reasonable approximation can be made that the filter corresponding to the cepstral mean (that is, the CCF) is minimum phase. A theoretical justification for this assumption is outlined in Appendix A.

The filter coefficients corresponding to the long-term cepstral mean can be evaluated using the reverse recursion of equation [3.5]. The expression for filter coefficients from the cepstral coefficients can be obtained by using the recursion [13] given by,

$$a_n = -c_n - \frac{1}{n}\sum_{k=1}^{n-1} c_{n-k}a_k \quad 1 \leq n \leq P. \tag{4.20}$$

The filter derived from the cepstral mean coefficients is termed as the **Channel Compensation Filter**. This filter (henceforth referred to as the CCF) represents an *inverse channel filter* that deconvolves the effect of the channel from the speech utterance.

In order to illustrate the effect of the CCF, a speech utterance from the TIMIT database, which was down-sampled to 8 KHz, was used. The duration of the utterance was approximately 10 seconds. The utterance was convolved with typical telephone channels obtained from the telephone channel simulator [88]. The CMV (Continental Mid Voice) channel and the CPV (Continental Poor Voice) channel were chosen. The frequency responses of the channles are shown in Figures (4.5) and (4.6).

The long-term cepstral mean was obtained by averaging 12^{th} order LP derived cepstral coefficients (LPCC) derived for each speech frame. The corresponding frequency response of the CCF evaluated by transforming the cepstral mean to filter coefficients is shown in Figure (4.7). The degrading channel was the CMV channel. A similar frequency response corresponding to the cepstral mean for speech degraded by the CPV telephone channel is shown in Figure (4.8).

One can observe from the frequency responses of the filters (CCFs), that they exhibit the characteristic response of a corresponding inverse channel. The inverse (or deconvolution) filter equalizes or effectively compensates for the bandpass effect of the convolutional distortion on each speech segment.

Let $N_{ccf}(z)$ represent the channel compensation filter (CCF) that corresponds to the all-pole approximation to the channel $N_{ch}(z) = \frac{1}{N_{ccf}(z)}$. A subtractive component in the cepstral domain corresponds to an FIR filter or a Moving Average (MA) component. Thus, one can represent the channel normalized spectrum of speech as,

[1] For a truncated cepstral sequence, one can assume it to be causal by padding infinite zeros to the sequence.

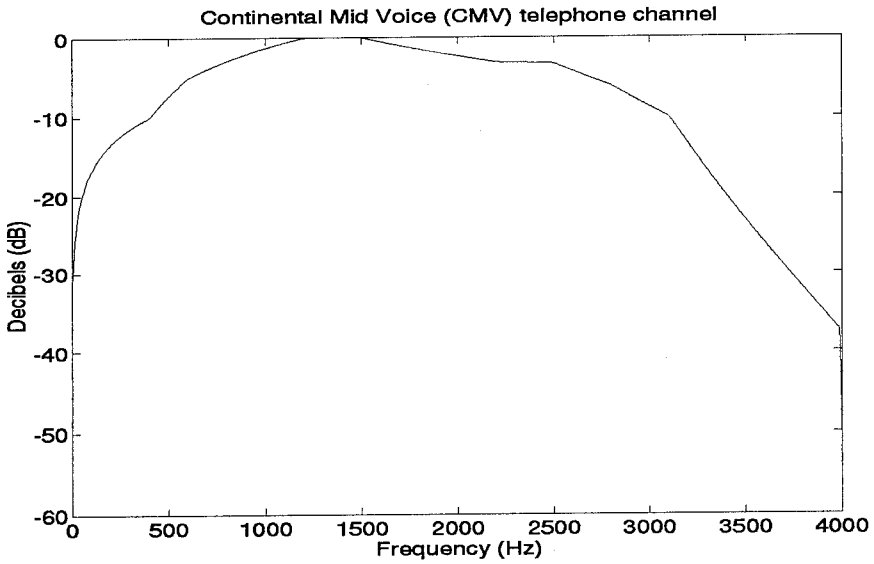


Figure 4.5: CMV Channel Response.

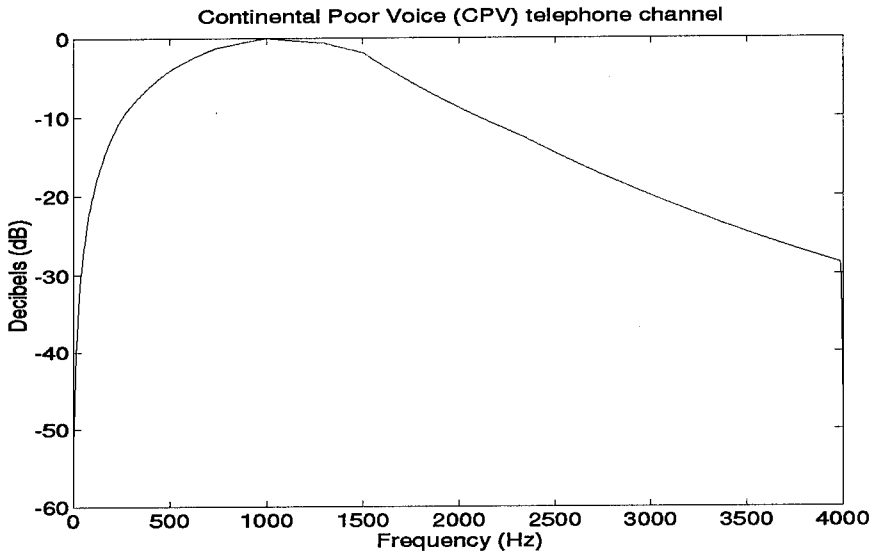


Figure 4.6: CPV Channel Response.

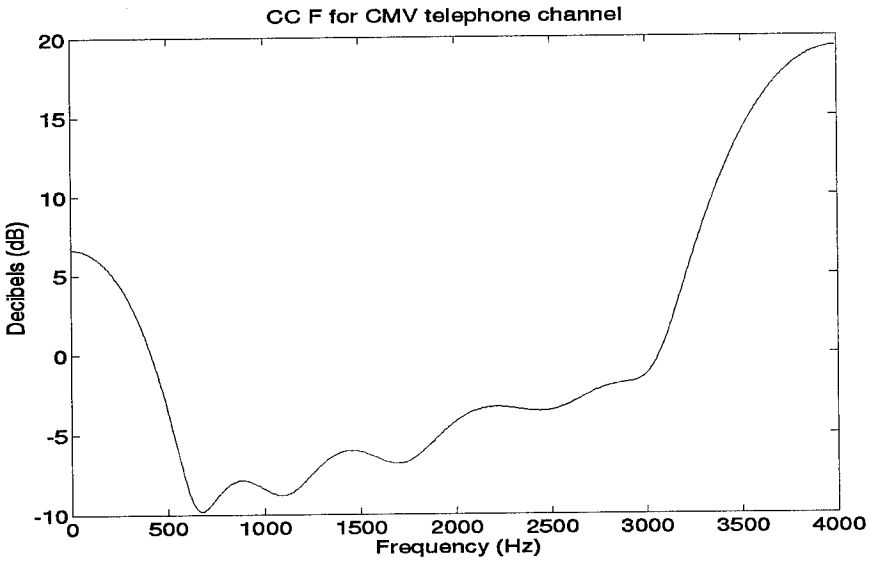


Figure 4.7: CMV inverse filter response.

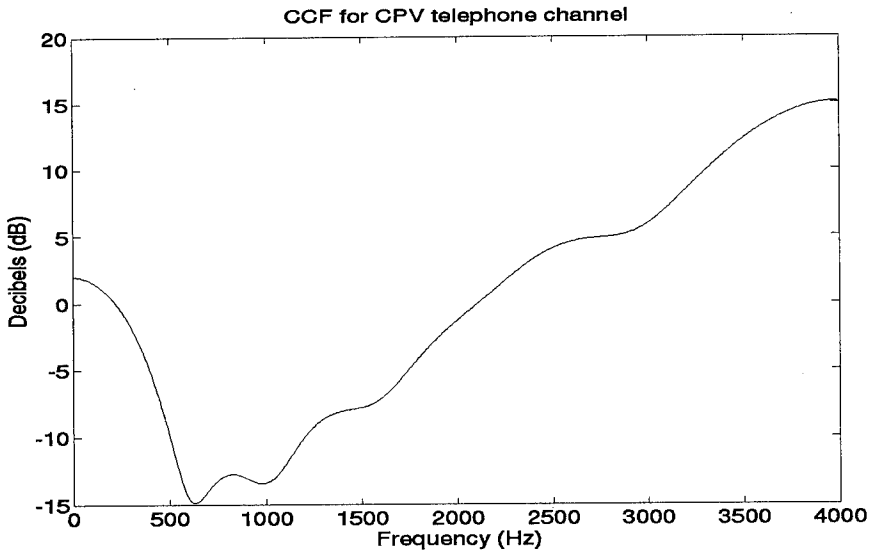


Figure 4.8: CPV inverse filter response.

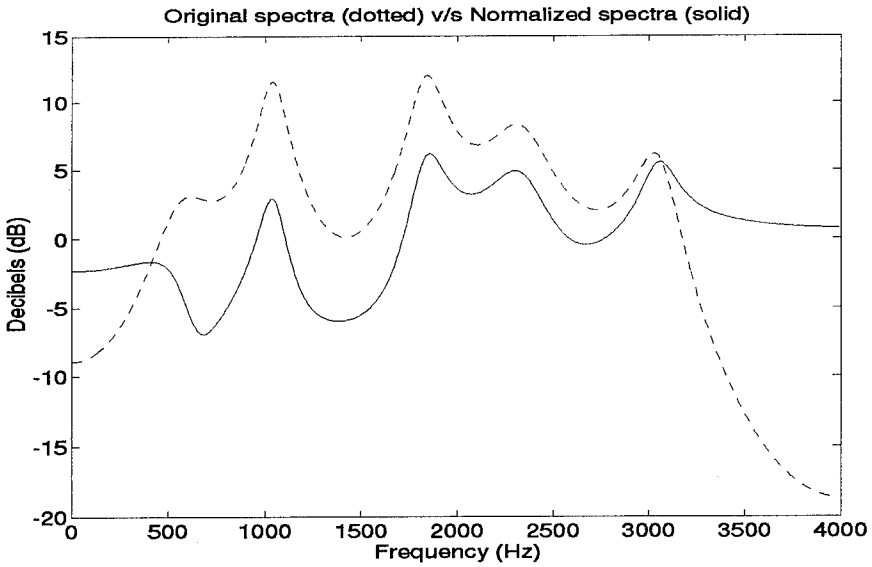


Figure 4.9: Effect of channel normalization for CMV channel.

$$Y_{normalized}(z) = \frac{N_{ccf}(z)}{A(z)}. \tag{4.21}$$

Figure (4.9), compares the spectrum of a voiced speech frame convolved with a CMV channel with the same spectrum of speech normalized by the estimated CCF from the cepstral mean of the utterance containing the same frame. A similar comparison for speech degraded by a CPV channel is shown in Figure (4.10).

Important deductions can be made by empirically observing the spectral content of the responses of the speech spectra. Due to the effect of the inverse filter, the spectral content in some frequency bands has been significantly attenuated. Such attenuation would typically occur for all speech frames from which the cepstral mean is being subtracted. This has a degrading effect on the accumulated spectral distortion over the entire speech utterance which is often calculated during classification.

It can also be observed that the estimate of the resulting inverse filter has a ripple in the passband when compared to a smooth passband in the actual channel responses.

Basis for Pole filtering

In this section, the effect of the CCF for a cross-channel scenario is investigated, wherein the speech utterance has been distorted by different transmission channels. A cross-channel scenario is most realistic wherein the convolutional distortions could be from different recording devices or different telephone channels for training and testing data.

In order to investigate this, the same speech utterance, convolved with a CMV channel and a CPV channel were investigated. It is apparent from Figures (4.5) and (4.6) that the attenuation of speech due to the CPV channel is much more pronounced, due to sharper roll-off at higher frequencies, leading to poorer voice quality.

In order to observe the effective channel normalization across channels, the normalization effect due to cepstral mean removal for both the channels is compared in Figure (4.11) for the same speech frame. One can observe that elimination of the cepstral bias does a reasonably good job in minimizing the mismatch between the two spectra.

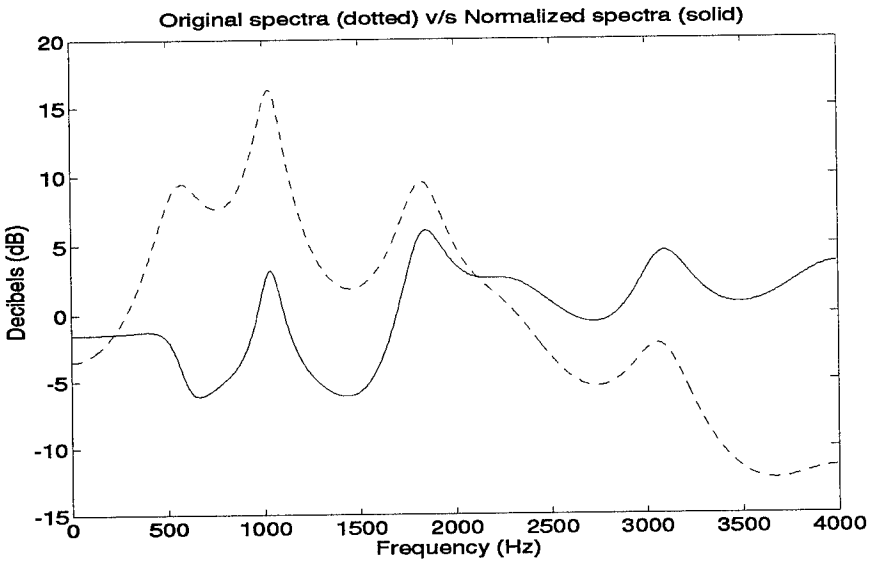


Figure 4.10: Effect of channel normalization for CPV channel.

However, the spectral information is attenuated in either case due to inaccurate estimates of the channel cepstrum, $\mathbf{c}_S$, biased by the invariant component due to speech, $\mathbf{s}_S$.

From the above discussion, one can make an important deduction. The estimate of the long-term cepstral mean, $\mathbf{c}_S$, can be improved so that,

- the spectral information in every frame of speech is preserved after channel normalization, while,
- the mismatch across channels is minimized.

Clearly, this can be achieved by de-emphasizing the contribution of the invariant component due to speech, $\mathbf{s}_S$, in the channel cepstrum, $\mathbf{c}_S$. In other words a methodology needs to be derived so that,

$$\mathbf{c}_S \rightarrow \mathbf{h}, \tag{4.22}$$

where $\mathbf{h}$, is the cepstral estimate of the actual convolutional distortion component.

A reasonably accurate estimate of the channel cepstrum, $\mathbf{c}_S$, could be obtained if the cepstral mean solely due to the clean speech, $\mathbf{s}_S$, with which the channels were convolved was available. In the case when the cepstrum due the clean speech is known, a reasonable accurate channel estimate,

$$\mathbf{h} \approx \mathbf{c}_S - \mathbf{s}_S, \tag{4.23}$$

can be evaluated. However, the cepstral mean of a speech utterance prior to convolution with the channel is never available in practice and hence it is impossible to entirely decouple the cepstral component due to speech from the cepstral component that corresponds to the channel.

From the observations made by studying the characterisitcs of the long-term cepstral mean and its spectral distribution for short utterances, the effects of the individual poles on the cepstral mean can be investigated. By studying the effect of the individual poles on the cepstral mean, algorithms can be developed that reduce the speech component in the cepstral mean and thereby improve the channel cepstrum estimate. The Pole filtering approach involves weighting and manipulating the poles of a speech frame in order to de-emphasize the component of the cepstral mean due to speech.

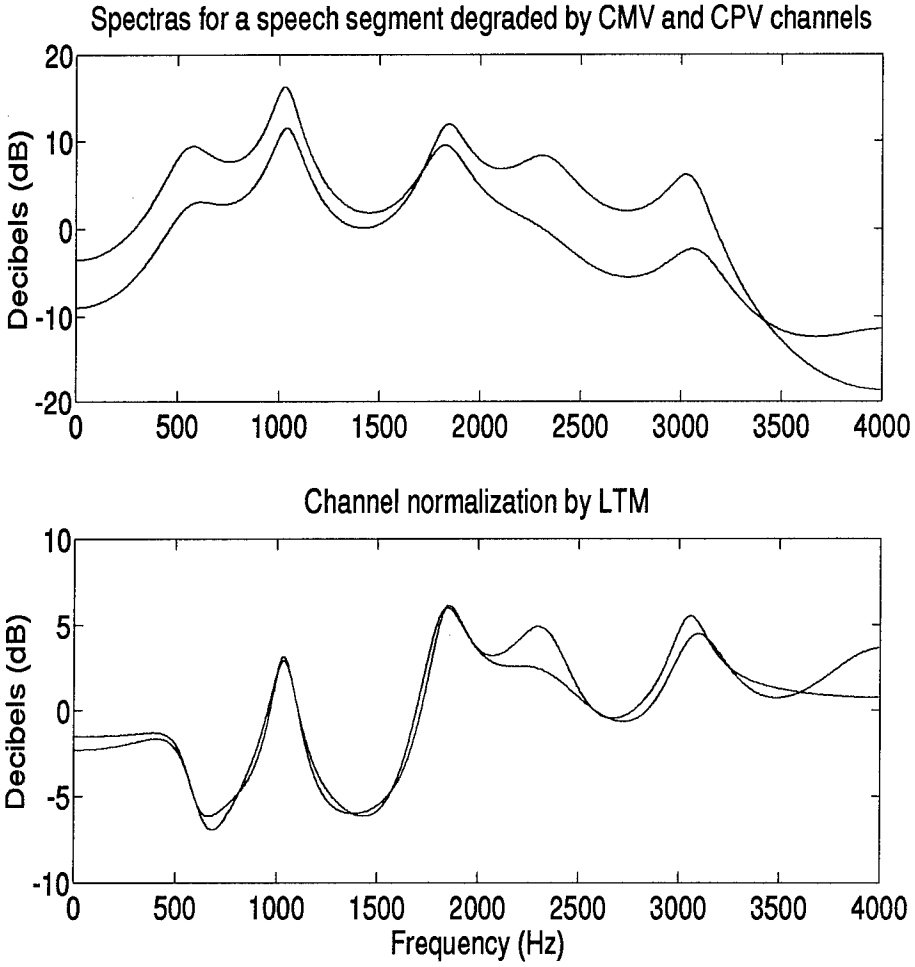


Figure 4.11: Effect of normalization implied by cepstral mean removal.

Approaches to Pole Filtering

The estimate of the channel cepstrum, $\mathbf{c}_S$, depends upon the number of speech frames available in the utterance. In the case where the speech utterance is sufficiently long, it is possible to get an estimate of the channel cepstrum that approximates the true channel estimate, $\mathbf{h}$. A key issue is the phonemic diversity of the utterance. In most practical situations, however, the utterance durations for training or testing are never long enough to allow for $\mathbf{s}_S \rightarrow 0$.

The cepstral mean estimate however can be improved by studying the dominance of the poles in the speech frame and their contribution to the estimate of the channel cepstrum.

The effect of each mode of the vocal tract on the cepstral mean can be investigated by decomposing the root power sum formulation to evaluate a *partial cepstral mean* due to each spectral component or a corresponding complex conjugate pole pair based on their dominance.

A spectral component, for a frame of speech, is most dominant if it corresponds to a complex conjugate pole pair closest to the unit circle (minimum bandwidth) and least dominant if it corresponds to a complex conjugate pole pair furthest from the unit circle (maximum bandwidth).

A simple experiment to study this effect was carried out by evaluating the partial long-term cepstral mean due to the most dominant spectral component (pole pair with the least bandwidth) in every frame of a speech utterance. The contribution to the long-term mean due to second most dominant spectral component was then evaluated and so on for the rest of the spectral components. This was carried out by sorting all complex conjugate pole-pairs (excluding real poles) of the all-pole LP filter according to their bandwidths in ascending order. Let $\langle z_{d_1}, z^{\star}_{d_1} \rangle$ be the most dominant complex pole-pair and $\langle z_{d_q}, z^{\star}_{d_q} \rangle$ for the q^{th} complex pole pair for poles $z_k, k = 1, \cdots, P$ for a frame of speech.

The partial cepstral mean $\mathbf{c}_{S_1}$ can be computed as,

$$\mathbf{c}_{S_1} = \left(\frac{1}{M} \sum_m \langle z^1_{d_1,m} + z^1_{d^{\star}_1,m} \rangle \quad \frac{1}{M} \sum_m \langle z^2_{d_1,m} + z^2_{d^{\star}_1,m} \rangle \quad \cdots \quad \frac{1}{M} \sum_m \langle z^P_{d_1,m} + z^P_{d^{\star}_1,m} \rangle \right). \tag{4.24}$$

The summation of partial long-term means, $\mathbf{c}_{S_1}, \mathbf{c}_{S_2}, \cdots, \mathbf{c}_{S_q}$, for all spectral components must be equal to the total long-term cepstral mean $\mathbf{c}_S$ after considering the spectral components corresponding to the real poles,

$$\mathbf{c}_S = \mathbf{c}_{S_1} + \mathbf{c}_{S_2} + \cdots + \mathbf{c}_{S_q}. \tag{4.25}$$

Figure (4.12) shows the partial cepstral means due to each dominant spectral component aggregated from all the frames of speech of an utterance. Individual CCFs were evaluated for each of the partial cepstral means to observe their spectral distribution. The individual responses are shown in Figure (4.13) for CMV channel and Figure (4.14) for the CPV channel.

One can observe from the individual responses of the CCFs, that the contribution to the partial cepstral mean due to the more dominant poles (or the narrow-band poles) is more biased by the spectral content relating to speech. In fact the inverse filter due to the narrow band poles exhibit characteristics that would attenuate, or *notch* useful spectral information when subtracted in the cepstral domain.

An improper estimate of the channel cepstrum causes a more drastic *nulling* effect on the spectra of a speech frame corresponding to an elimination of the channel cepstrum from the cepstrum of the frame of speech. One can also observe the zeros of the CCF for the CMV and the CPV channels on the unit circle. It can be seen that certain zeros of the inverse channel estimate (or poles of the channel estimate) obtained are relatively close to the unit circle, unlike the zeros of a real channel inverse which would be expected to be composed of generally broad-band spectral components.

The zeros of the inverse channel estimates obtained from the cepstral mean for speech degraded by the CMV channel and the CPV channel are shown in Figures (4.15) and (4.16).

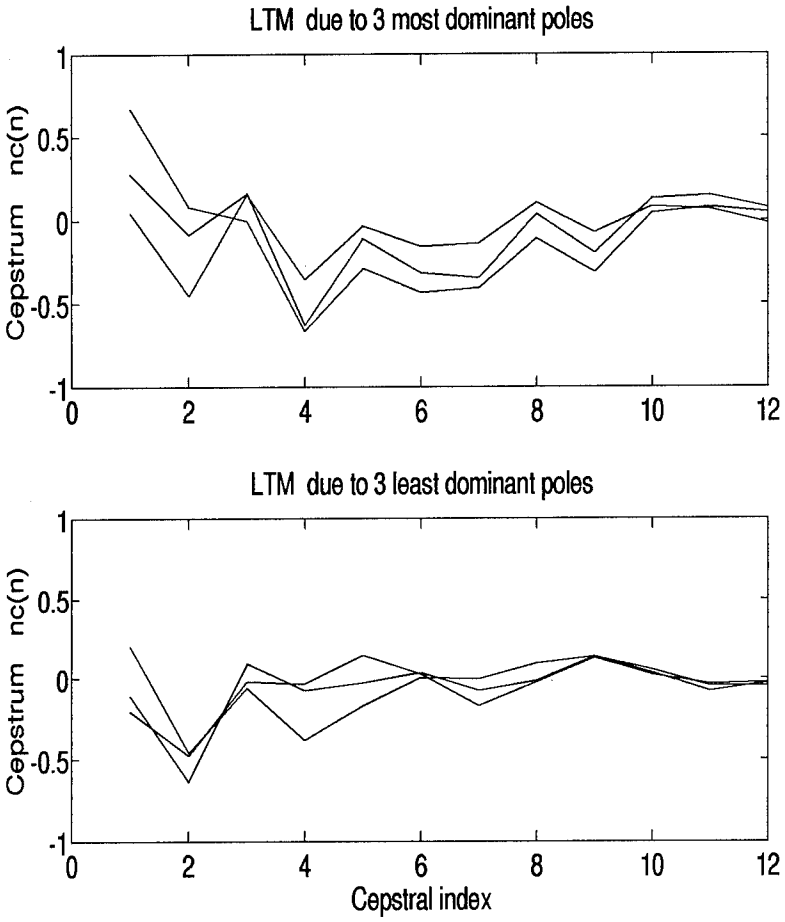


Figure 4.12: Partial cepstral means for speech degraded by CMV channel.

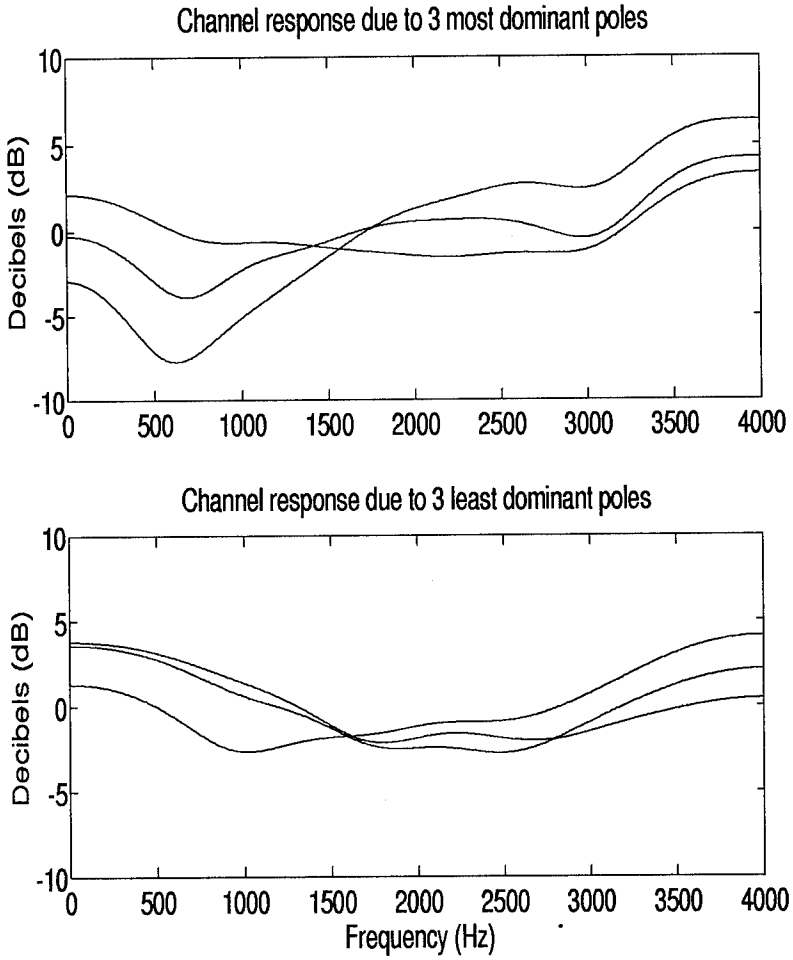


Figure 4.13: Responses for partial cepstral means for speech degraded by CMV channel.

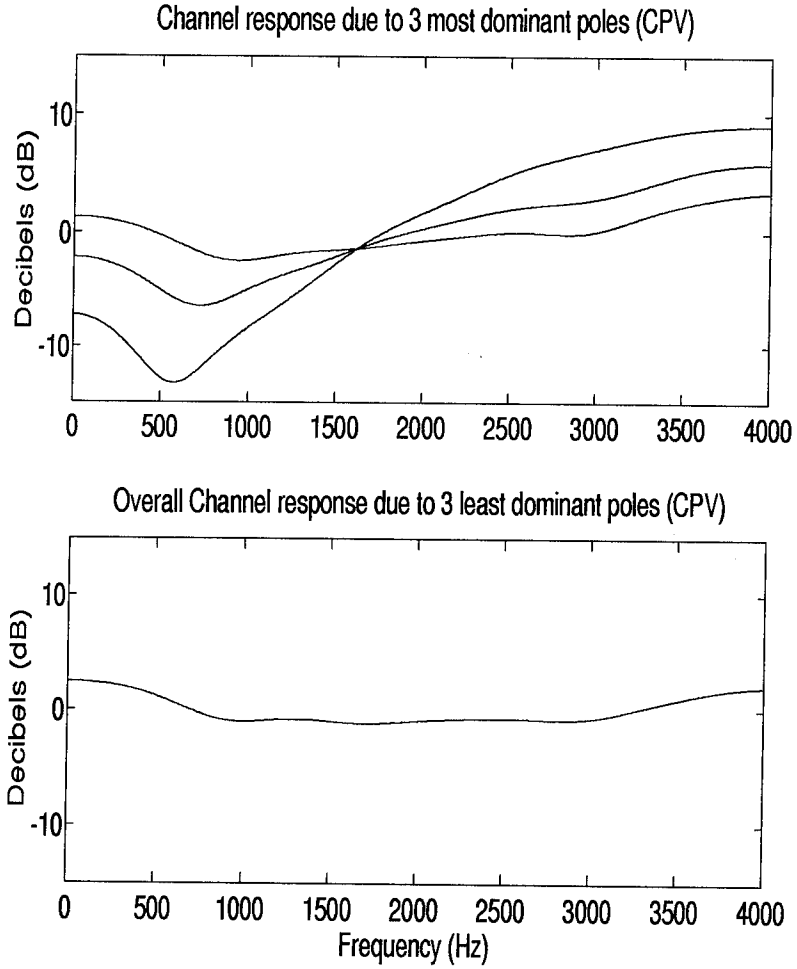


Figure 4.14: Responses for partial cepstral means for speech degraded by CPV channel.

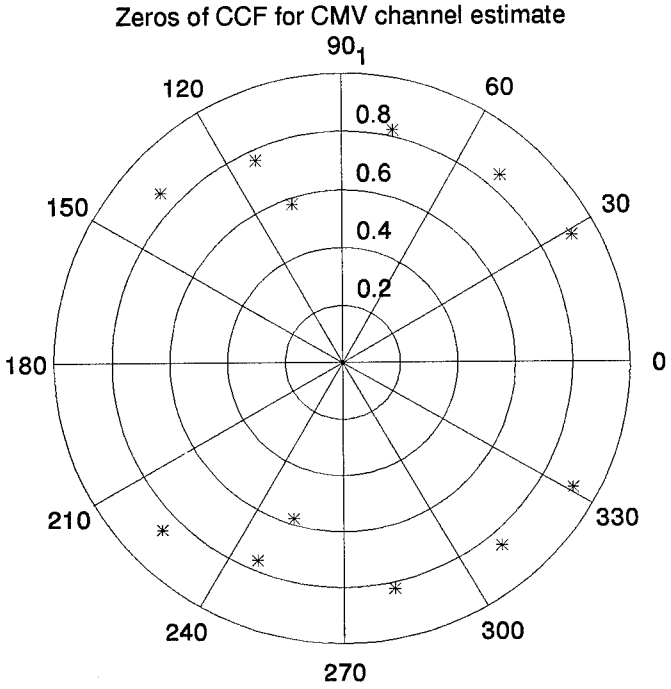


Figure 4.15: Channel zeros estimated from cepstral mean for CMV channel speech.

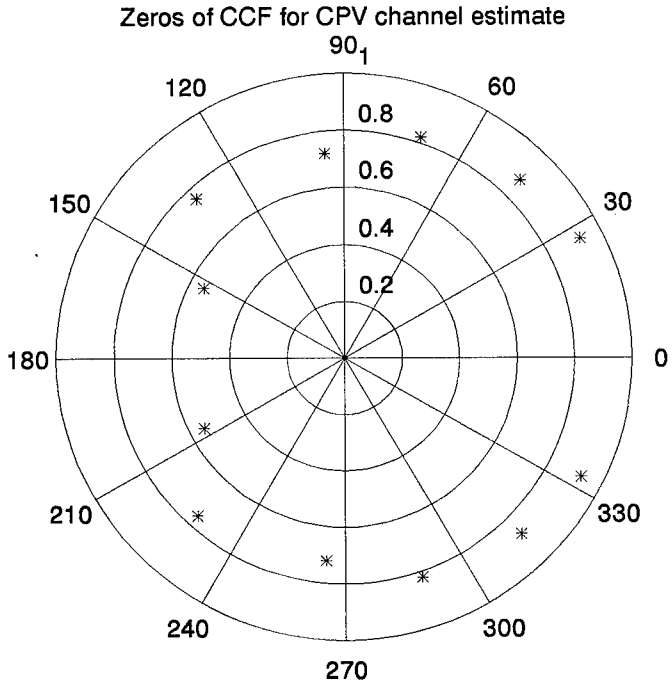


Figure 4.16: Channel zeros estimated from cepstral mean for CPV channel speech.

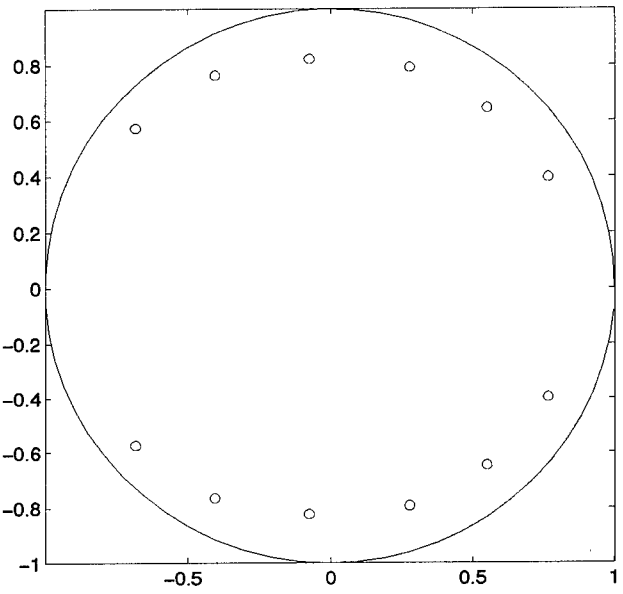


Figure 4.17: Channel poles estimated from the impulse response of CMV channel.

The contribution to the long-term cepstral mean by the broad-bandwidth poles tends to exhibit smoother inverse filtering characteristics. Also, the estimate of the inverse channel will not attenuate much spectral information when subtracted in the cepstral domain.

The analysis of the partial cepstral means lends insight into improving the channel cepstrum estimate. One could empirically observe that the poles of an all-pole filter approximating am actual channel response comprise of sufficiently broad bandwidth components, and hence implies a smoother bandpass filter (observe Figures (4.5) and (4.6)). This assumption can be justified by fitting an all-pole model to the impulse response of the two known channel distortions (CPV and CMV) obtained from the simulator. The all-pole LP spectra of the impulse responses are shown on the unit circle in figures (4.17) and (4.18). One can observe that the poles approximate a broad band channel response and exist further away from the unit circle. The channel estimate generally does not consist of poles close to the unit circle which is often signified by high-Q regions in the spectra.

Constraining the poles of speech in order acquire a smoother and hence a more accurate inverse channel estimate in the cepstral domain, corresponds to a modified cepstral mean, $\mathbf{c}_S^{pf}$ that de-emphasizes the cepstral bias related to the invariant component due to the speech. The refined cepstral mean removal, devoid of the gross spectral distribution component due to speech offers an improved channel normalization scheme.

Techniques have been proposed in the subsequent section to improve the channel estimate by intelligent manipulation of dominant modes in the speech frame.

Channel normalization based on Pole filter Cepstral Coefficients

Pole filtering using Selective Pole Manipulation

One technique of improving the estimate of the channel uses Pole filtered cepstral coefficients (PFCC) wherein, the narrow band poles are inflated in their bandwidths while their frequencies are left unchanged. The effect is equivalent to moving the narrow band poles inside the unit circle along the same radius, thus keeping the frequency constant while broadening the bandwidths. The procedure has been illustrated in Figure (4.19).

PFCCs are evaluated for every speech frame concurrently with the LPCC, the only difference being that if a pole in the speech frame has a bandwidth less than a pre-determined threshold (α), the bandwidth

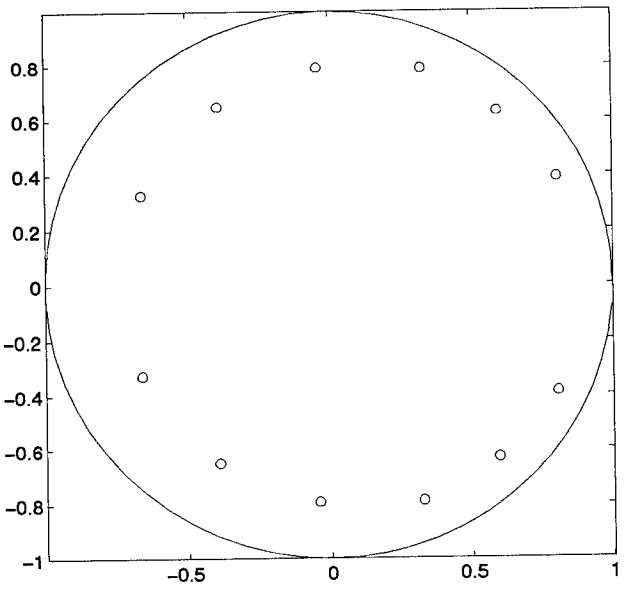


Figure 4.18: Channel poles estimated from the impulse response of CPV channel.

of that pole is clipped to that threshold. The algorithm to calculate the PFCCs is listed below.

For each Speech frame:

Evaluate the roots z_k of the LP polynomial.
if $|z_k| \geq \alpha$, (α=pole bandwidth threshold),
$|z_k| = \alpha$
Modify z_k to $\tilde{z_k}$ by,
$\tilde{z_k} = \alpha \angle z_k$;
endif
Evaluate **LPCC** using z_k.
Evaluate **PFCC** using $\tilde{z_k}$.

The PFCCs are used to evaluate a modified long-term cepstral mean, $\mathbf{c}_S^{pf}$. The CCF characteristics obtained from averaging the PFCCs when the narrow band pole with radius greater than $\alpha = 0.9$ is thresholded to 0.9 is shown in Figure (4.20).

An improved inverse filter estimate is obtained by using the mean of PFCCs which better approximates the true inverse channel filter. The modified cepstral mean, when subtracted from speech cepstra of individual speech frames tends to preserve the spectral information while more accurately compensating for the spectral tilt of the channel.

A typical effect of channel normalization using the modified cepstral mean subtraction obtained from averaging the PFCCs on a voiced frame of speech is shown in Figure (4.21). One can see that substantial spectral information is preserved with the new channel cepstrum estimate especially in the lower and higher frequency bands.

It should be noted that the PFCCs are used only to estimate the channel cepstrum i.e. the long-term cepstral mean. The modified long-term mean obtained from the PFCCs is subtracted from the LP derived cepstrum of every speech frame of the utterance instead of a conventional cepstral mean. The processing algorithm is outlined below.

ALGORITHM:

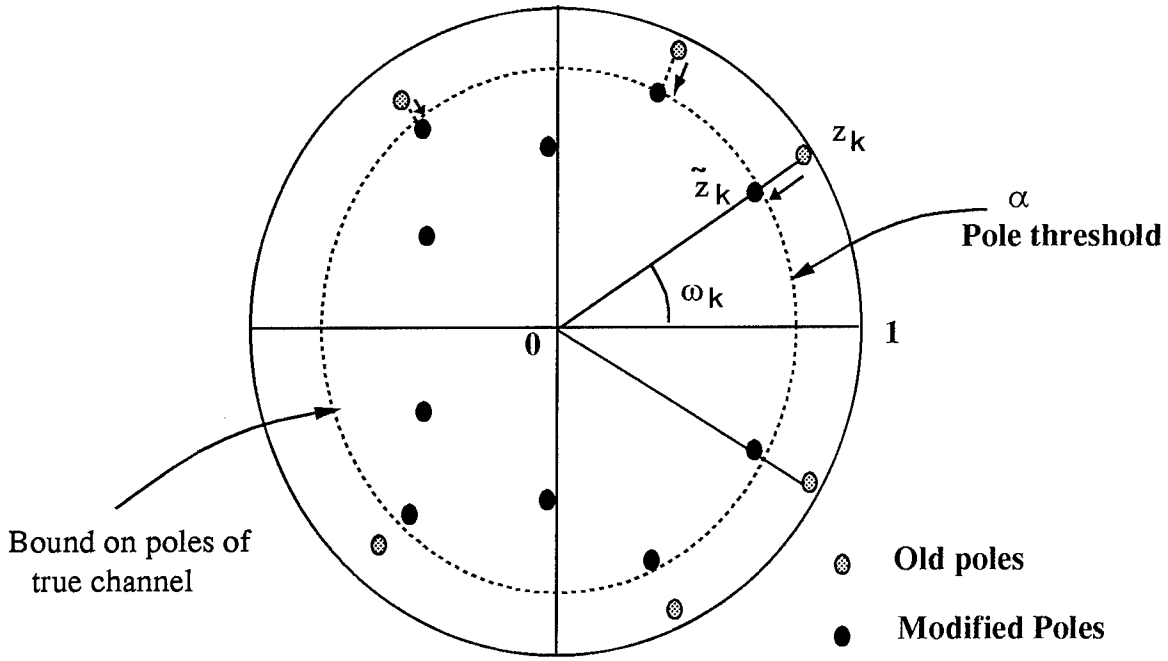


Figure 4.19: Pole thresholding process on the unit circle.

- *Evaluate long-term cepstral mean based on PFCCs.*
- *Normalize LPCCs by subtracting a refined cepstral mean obtained from averaging the PFCCs.*

The estimate of the modified cepstral means using PFCCs and their corresponding CCFs for different bandwidth thresholds is shown in the Figure (4.22).

The justification for the choice of pole bandwidth thresholds, α, can be made from the observing the all-pole fit to the true channel impulse responses or the zeros of the actual inverse channel estimate in Figures (4.17) and (4.18). One can observe that the poles of the channel are sufficiently broad-band compared to the dominant pole in a typical voiced speech frame. In fact a smooth channel characteristics for a given all-pole filter order would consist of only broad-band poles.

Pole filtering using weighted prediction coefficients

Broadening of bandwidths of pole can also be achieved by weighting the prediction coefficients to evaluate the spectrum by,

$$A(\gamma z) = 1 + \sum_{k=1}^{P} a_k (\gamma z)^{-k}, \tag{4.26}$$

and the corresponding cepstral transformation is,

$$c_S^{pf}(n) = \gamma^n c(n), \tag{4.27}$$

where γ with a value between 0 and 1, is a pole bandwidth broadening factor. Such bandwidth broadening has been used in speech coding for perceptual shaping of the quantization noise [93] and for improving the distortion measures in the presence of noise [75]. The value of $\gamma = e^{-(\pi \frac{\delta}{f_s})}$, is based on δ Hz, which is the frequency with which the pole bandwidths can be broadened. The pole-filtered cepstral mean obtained after weighting the cepstral coefficients with decaying factor, is subtracted from LP cepstra for every frame of the speech utterance.

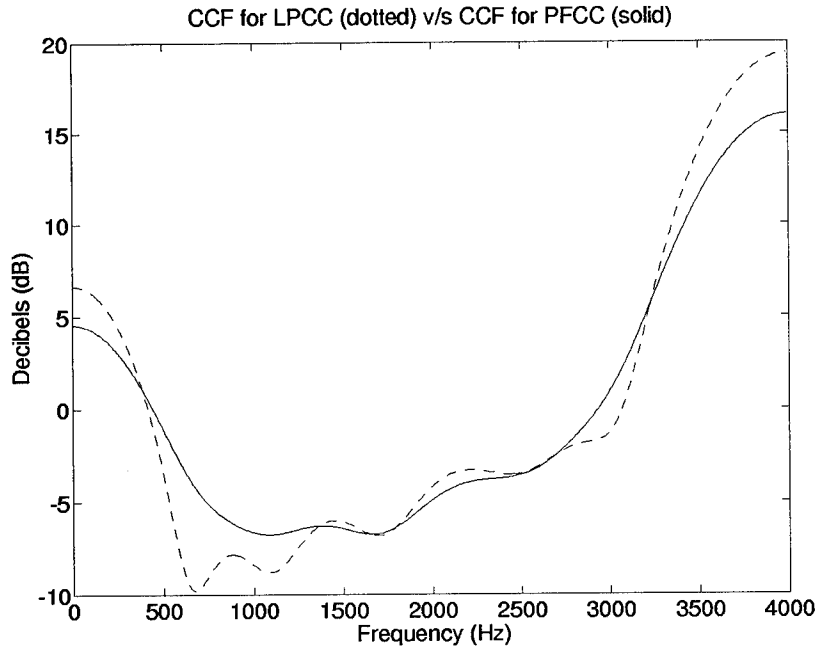


Figure 4.20: Improved channel estimate using cepstral mean of PFCC.

The disadvantage of this method is that it broadens the bandwidths of all the poles modeling the speech frame. Thus the broad band poles which may be critical in estimating the roll-off due to the convolutional distortion, also are broadened in bandwidth. The pole-filtered cepstral mean estimates obtained using this approach are not as effective in channel normalization for improved recognition accuracy as those obtained via selective pole manipulation.

The interframe pole-filtering approach outlines a technique wherein poles parameters are selectively modified to improve the channel cepstrum estimate. The pole-filtered cepstral mean subtraction technique implies a filtering scheme that improves the robustness of cepstral features in the presence of channel differences. In Chapter Five results are reported using various emperically decided pole bandwidth thresholds. Simulations will be presented that show an improvement in the recognition performance by using a modified cepstral mean estimate for channel normalization.

In the subsequent section a previously proposed adaptive spectral component weighting intraframe processing approach is explained in terms of the Pole-filtering paradigm.

4.3.2 Intraframe Pole-filtering approach

The intraframe approach consisting of an adaptive weighting of the spectral components to minimize channel mismatch within each individual speech frame was proposed previously by Assaleh and Mammone [31]. The technique can be considered as a complementary approach to the interframe pole-filtering approach and will be reviewed here.

The Adaptive Component Weighting (ACW) approach involves an intraframe pole weighting scheme which emphasizes the parts of the spectra which correspond to narrowband components due to speech and attenuates the more channel sensitive broadband components. Note that the complementary nature of the interframe and the intraframe approach is clear since the narrow bandwidth components that were de-emphasized in the interframe approach which yield improved channel estimates, are now emphasized in the intraframe approach. It was shown in the experiment discussed in Section (4.3) [31], that the LP

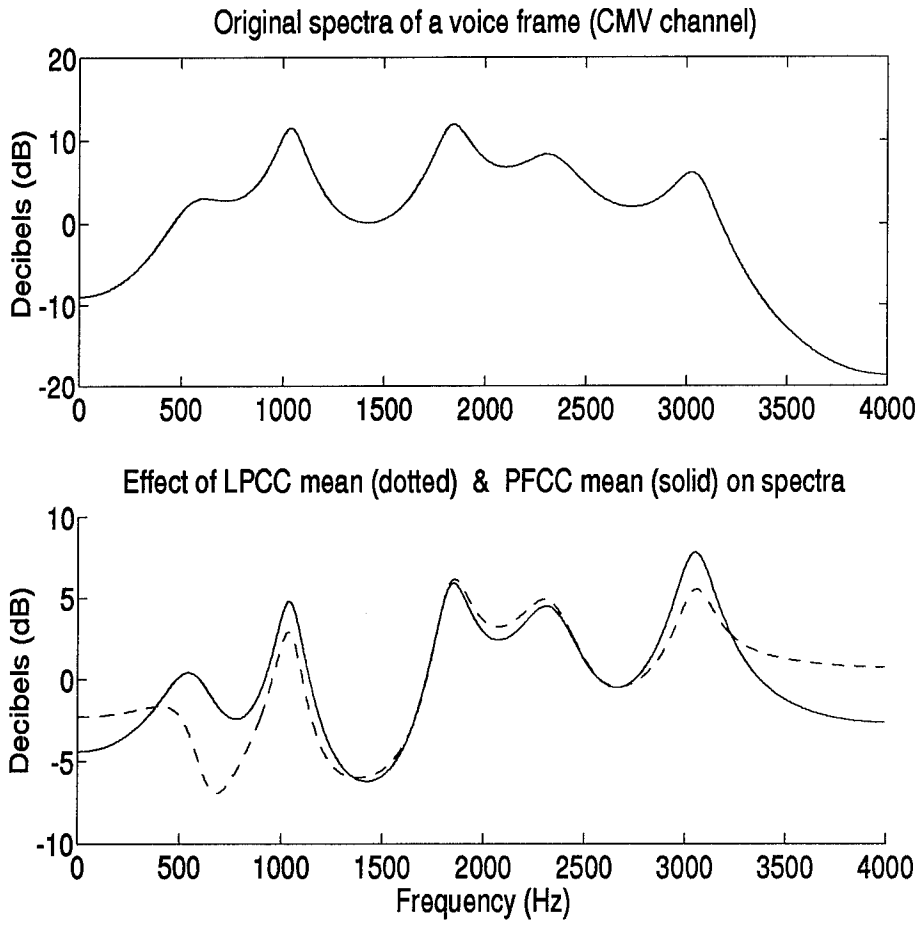


Figure 4.21: Channel normalization using ordinary cepstral mean v/s pole filtered cepstral mean.

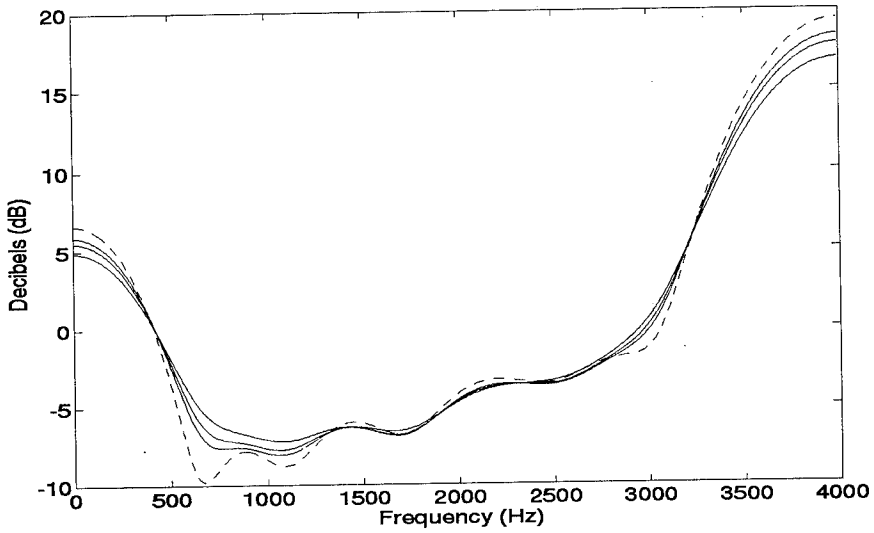


Figure 4.22: CCF for PFCC mean for different pole bandwidth thresholds at $|z|$=0.9,0.88,0.86 compared to ordinary LPCC mean (dotted).

spectrum should be modified so as to eliminate the spectral parameters that are more sensitive to channel variations. The ACW approach is based on modifying the LP spectra to minimize the channel mismatch between training and testing in cross channel conditions.

The ACW approach normalizes the residues r_k (which show the largest amount of variation with respect to channel differences), by setting $r_k = 1$ in equation [4.8], which can be viewed as weighting the k^{th} component by $\frac{1}{r_k}$. This normalization results in a modified spectrum which is referred to as the ACW spectrum. The ACW spectrum is given by

$$\hat{S}(z) = \sum_{k=1}^{P} \frac{1}{(1 - z_k z^{-1})} = \frac{N(z)}{1 + \sum_{k=1}^{P} a_k z^{-1}}, \tag{4.28}$$

where

$$N(z) = \sum_{i=1}^{P} \prod_{k=1 \neq i}^{P} (1 - z_k z^{-1}), \tag{4.29}$$

which can be rewritten in the form

$$N(z) = P(1 + \sum_{k=1}^{P-1} b_k z^{-k}). \tag{4.30}$$

Thus the normalization to the LP spectrum modifies each spectral component to yield a peak value of

$$\frac{1}{(1 - z_k z^{-1})}|_{z=e^{j\omega_k}} = \frac{1}{1 - |z_k.|} \approx \frac{1}{B_k}. \tag{4.31}$$

Equation [4.31] shows that the ACW spectrum emphasizes the formant structure by weighting each component by $\frac{1}{B_k}$. Thus narrow-bandwidth components are amplified and broad-bandwidth components are attenuated.

$\hat{S}(z)$ is no longer an all pole autoregressive (AR) transfer function, as it now has a MA filter represented by $(P-1)$ zeros. This MA filter introduced by normalizing the residues can be viewed as an FIR filter.

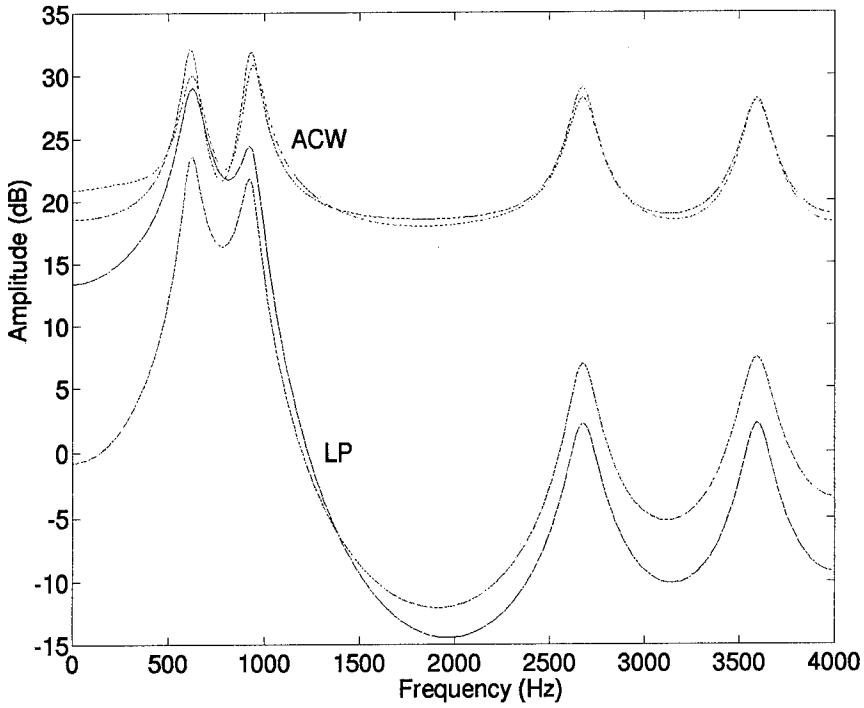


Figure 4.23: The channel effect on the composite LP and ACW spectra (adapted from Assaleh and Mammone [31]).

This filter creates a spectrum wherein the peak values of each component are inversely proportional to their bandwidths.

The channel effect on the composite LP and ACW spectra is shown in figure 4.23. It is obvious that the mismatch between the LP spectra before and after processing through the channel is much larger than that between the corresponding ACW spectra.

In the cepstral domain, the introduction of $N(z)$ results in a subtractive component to the all-pole cepstrum. This component varies with each frame unlike the common intraframe processing techniques that apply a set of fixed weights to the all-pole cepstrum, $c(n)$.

The subtractive cepstral component, $c_N(n)$, which is associated with $N(z)$, can be obtained by its recursive relation with filter coefficients b_k in equation [3.5].

Thus the ACW cepstrum, $c^{acw}(n)$ can be obtained as follows:

$$c^{acw}(n) = c(n) - c_N(n). \tag{4.32}$$

This method of intraframe pole filtering was shown to yield substantial improvements under cross channel scenarios.

ACW cepstral mean as a pole-filter estimate of the ordinary cepstral mean

It was also shown that ACW intraframe weighting followed by ACW cepstral mean removal further improved the recognition accuracy [31]. One can prove that the improvement in the performance due to ACW mean removal may be due to the fact the the ACW mean removal itself, implicitly corresponds to eliminating another form of a pole-filtered cepstral mean estimate. Averaging the adaptive subtractive cepstral component $c_N(n)$ for every frame corresponds to a pole-filtered cepstral mean estimate given by,

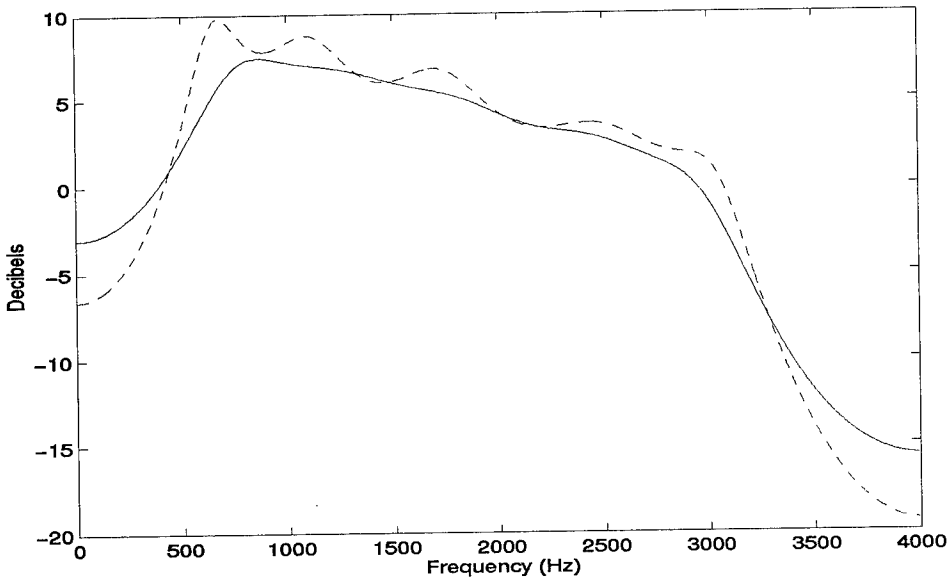


Figure 4.24: Comparison of spectra of ordinary cepstral mean (dotted) with ACW subtractive cesptral component mean (solid) for CMV channel speech.

$$\mathbf{c}_{acw}^{pf} = \left(\langle \textstyle\sum_{m=1}^{M} c_N(1) \rangle \quad \langle \textstyle\sum_{m=1}^{M} c_N(2) \rangle \quad \cdots \quad \langle \textstyle\sum_{m=1}^{M} c_N(Q) \rangle \right), \tag{4.33}$$

where Q is the order of the cepstrum and M is the total number of frames in the utterance.

The frequency responses corresponding to the cepstral mean of the subtractive ACW component for speech degraded by CMV and CPV channels is shown in Figures (4.24) and (4.25). One can observe that the response curves have similar smooth inverse channel characteristics that were obtained using pole-filtering.

4.3.3 Combined Interframe and Intraframe pole-filtering

Channel normalization methods in the past have often relied on combining conventional interframe and intraframe processing. For many intraframe processing techniques, interframe processing such as CMN have been implicitly assumed. Long-term mean removal following an intraframe cepstral weighting has been found to help normalize the channel mismatch.

An improved approach proposed here estimates the inverse channel filter using interframe pole-filtering and equalizes the effect of the channel on a speech utterance by deconvolution in the time domain. The deconvolution filter is obtained by converting a pole-filtered cepstral mean estimate to the corresponding CCF. A conventional intraframe weighting followed by cepstral mean removal is then carried out on a second pass to equalize the residual channel mismatch. The two-pass approach is compared with conventional channel normalization approach in Figure (4.26). Although computationally expensive, the two pass approach is shown to perform better in minimizing the channel mismatch effects. The results are reported in Chapter Five.

The advantage of a two-pass channel normalization approach is illustrated in Figures (4.27) and (4.28). Figure (4.27) compares mismatch in speech spectra degraded by two channel CMV and CPV before and after deconvolving the estimated channel. Figure (4.28) compares the same for the ACW spectra. In either case one can observe that the spectral mismatch is minimized by deconvolving the inverse channel estimate,

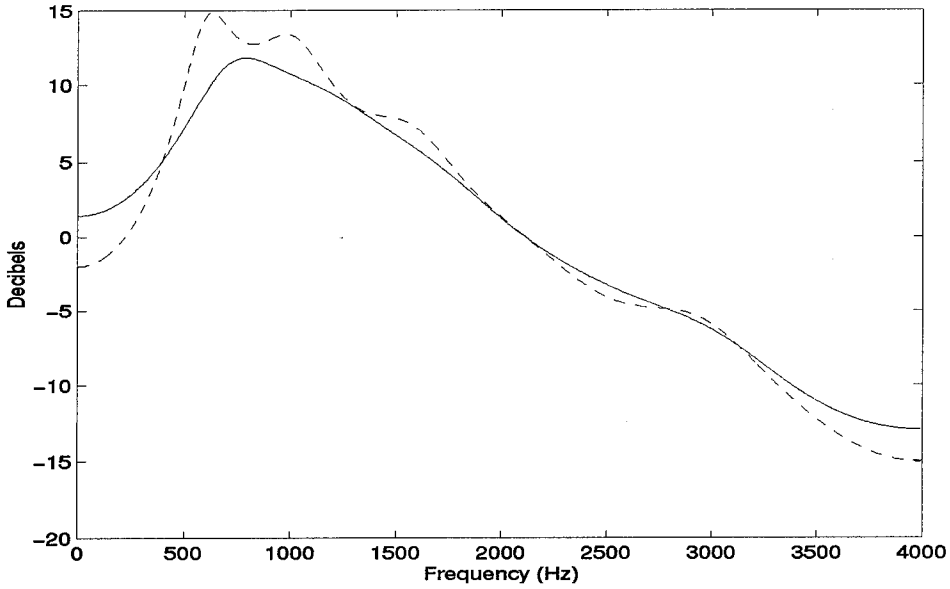


Figure 4.25: Comparison of spectra ordinary cepstral mean (dotted) with ACW subtractive cepstral component mean (solid) for CPV channel speech.

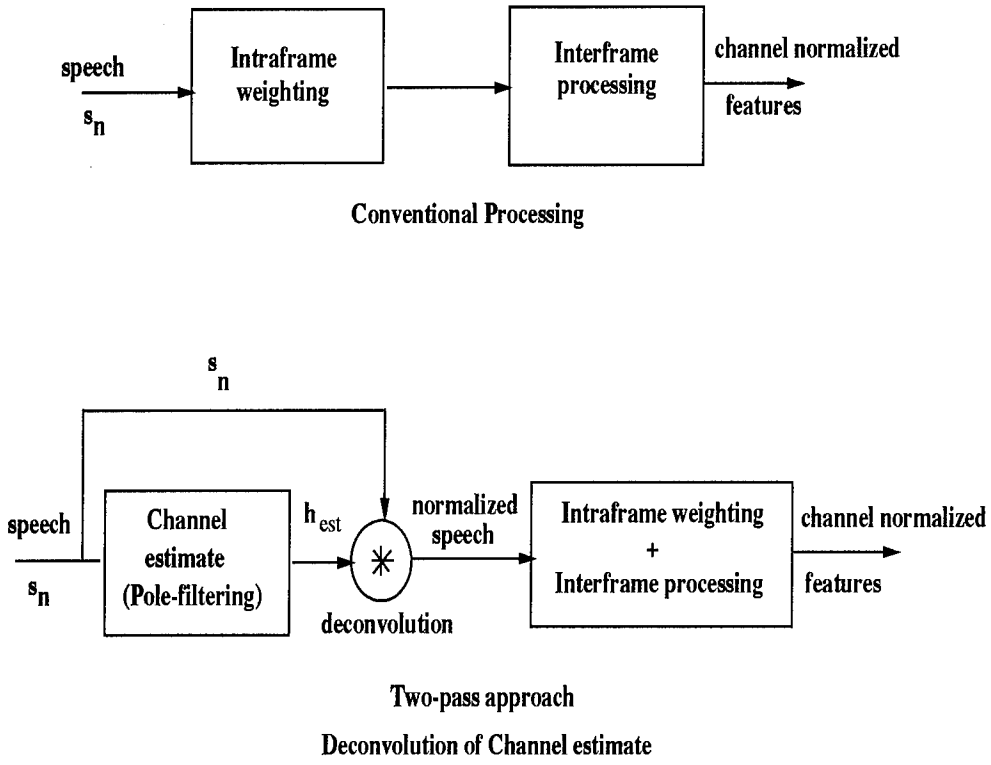


Figure 4.26: Combining inter-frame and intra-frame processing.

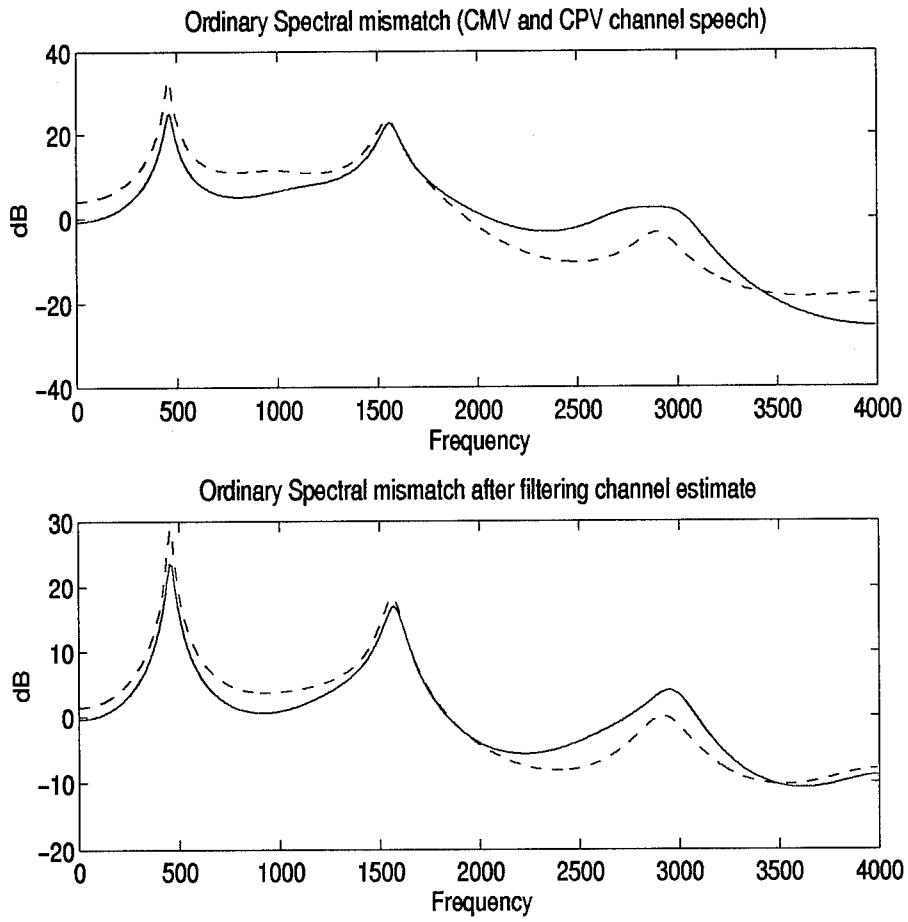


Figure 4.27: Spectral mismatch for a frame of speech convolved with CMV (solid) and CPV channel (dotted), top:one-pass, bottom: two-pass.

obtained from the pole-filtered cepstral mean, from the speech frame prior to the intraframe ACW based cepstral weighting approach.

The following chapter presents the experimental results on various benchmark databases for closed set text-independent speaker identification. The improved performance obtained using pole-filtered cepstral mean removal are reported. The improvements in recognition accuracy obtained by using a two-pass channel normalization approach over one-pass approach are also reported.

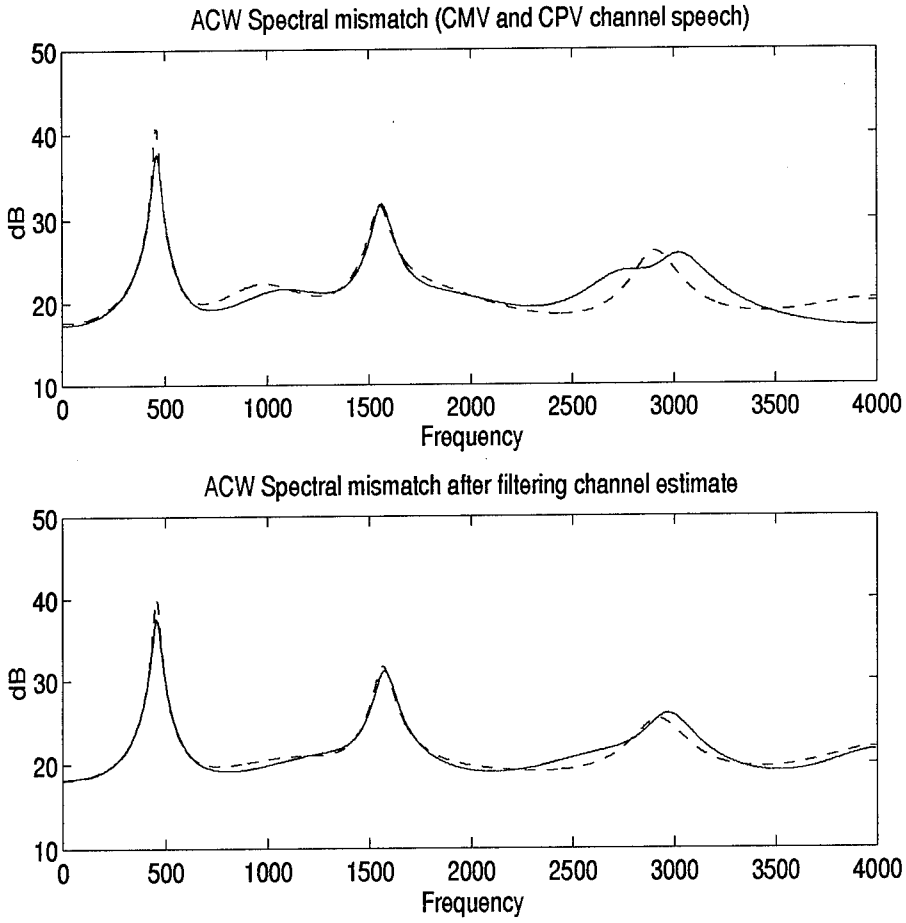


Figure 4.28: ACW Spectral mismatch for a frame of speech convolved with CMV (solid) and CPV channel (dotted) , top:one-pass, bottom:two-pass.

Chapter 5

Experimental Results

This chapter presents several experiments on channel normalization for closed set, text independent speaker identification systems. The improvements obtained using the pole filtering approach are tabulated and compared with conventional methods. The experiments have been presented in two parts. In the first part, the experiments were performed on simulated telephone channels obtained from a Wireline simulator [88]. The clean speech from the TIMIT [90] database was bandlimited and processed through the Wireline similator to generate telephone quality speech. The second part of channel normalization experiments were conducted on Benchmark corpuses containing real telephone conversations. The corpora were chosen from, the NTIMIT [91] (telephone quality TIMIT) and the KING database [89]. The results have been compared to conventional CMN which was proven to yield superior recognition performance compared to other interframe cepstral feature processing techniques [31, 64], for text independent speaker identification. In all the experiments the emphasis has been on short training and testing utterances. The experiments emphasize two critical issues in speaker recognition,

- Availability of limited data, and,
- Poor transmission channel conditions.

The TIMIT and the NTIMIT databases have been used to illustrate the limited data availability, while the KING experiments emphasize performance issues under severe channel mismatches among training and testing. For training and testing across channels it was found that the ordinary cepstrum (LPCC) performance was extremely poor and hence the improvements in the recognition accuracy has been reported by considering the performance using cepstral mean subtraction as the baseline for comparison.

The speech utterances were processed in overlapping speech segments of 30 msec, with a 10 msec update. The frames were passed through a pre-emphasis filter $1-0.95z^{-1}$ and windowed using a Hamming window. The order of Linear prediction was kept constant ($P = 12$) in all experiments. The order of the cepstral coefficients was also fixed ($Q = 12$).

The non-parametric K–means or VQ classifier was used for classification [92]. Training consisted of building a codebook of 46 codewords to model each speaker. The choice of 46 codewords roughly relates to number of phonemes spanning the phonetic space. The test utterance corresponding to the unknown speaker was processed to form test vectors. A Euclidean distance score was computed between the test vectors and each of the codebooks. The codewords which were closest to the test vectors were evaluated and the distances stored. The codebook associated with the minimum accumulated distance was classified as the speaker associated with that codebook.

The following abbreviations have been used in the tabulated results,

- LPCC-MR is ordinary cepstral mean subtraction (section 3.2.2).

- PFCC-MR is pole-filtered cepstral mean subtraction (section 4.3.1).
- LPCC-IC is cepstral feature extraction after deconvolving the actual inverse channel.
- FS corresponds to pole based frame selection [31].
- PFCC-MR(ACW) is the pole-filtered cepstral mean estimate obtained by averaging the subtractive component from the ACW approach (section 4.3.2).
- α is the pole bandwidth threshold, $|z|$.

The choice of the pole bandwidth threshold, α, was chosen where $\alpha \in [0.85, 0.9]$. The justification for the choice of this range of bandwidth threshold was outlined in section 4.3.1. The experimental results have been reported below along with an explanation of training and testing conditions on the benchmark databases.

5.1 Experiments on TIMIT and NTIMIT database

The TIMIT [90] database consists of 10 sentences spoken by each of the 630 speakers from 8 major dialect regions. There are 438 male and 192 female speakers distributed among the dialect regions. The TIMIT corpus is divided in the two sets for training and testing. The experiments were performed by choosing speakers from the "train" section of the database. Out of the 10 sentences spoken by each speaker, 5 sentences are labeled SX, 3 SI and 2 SA sentences. For all the experiments the training was carried out on the SX sentences and testing was carried out on the SA and the SI sentences (individual or concatenated). The training sentence durations roughly corresponded to about ten seconds of spoken material for every speaker. Each of the SA and SI sentences used for testing varied from 0.7 seconds to 3 seconds in duration after speech/non-speech discrimination.

The NTIMIT (Network TIMIT) [91] corpus consists of the TIMIT database transmitted through a telephone network. The transmission involved the use of a commercial device to simulate the acoustic characteristics between a human mouth and a handset. Thus speech in the NTIMIT represents real telephone speech transmitted after being acquired via a carbon-button telephone handset.

5.1.1 Simulated channel experiments

Experiments on Telephone channels obtained from the Wireline simulator [88] were performed using clean speech obtained from the TIMIT database. The main purpose of using a simulator was to study the effect of CMS and pole-filtered CMS for when training and testing were on the same channel and across channels. The use of a known convolutional distortion also helped in establishing and upper bound on the ideal recognition performance when the perfect inverse channel estimate were available for Channel normalization. Simulated CMV and CPV channels discussed in Chapter Four were chosen since they represent a significant channel mismatch.

The experimental conditions consisted of compiling a 38 speaker training and testing set from a section of the TIMIT database. The speakers were chosen from the New England dialect. Of the 10 SA, SI and SX sentences, the 5 SX sentences were concatenated and used for training while each individual sentence from the SA and SI was used for testing. The speech was first down sampled from 16 KHz to 8 KHz to simulate telephone bandwidth. The CMV and the CPV channels were used to simulate telephone channel conditions wherein the downsampled speech was filtered through the telephone simulator for degradation by either the CMV or CPV channels. The pole-filtered cepstral coefficients were evaluated by using a pole bandwidth threshold on $|z| = 0.86$ on the unit circle. The silence removal was carried out by a pre-determined energy threshold. All feature calculations were carried out after silence removal.

Method	Training	Testing	Accuracy(%)
LPCC-MR	CMV	CMV	63.1
PFCC-MR	CMV	CMV	69.5
LPCC-IC	CMV	CMV	86.8
LPCC-MR	CPV	CPV	62.1
PFCC-MR	CPV	CPV	68.9
LPCC-IC	CPV	CPV	85.7

Table 5.1: Single channel experiment on 38 speaker subset of TIMIT using pole-bandwidth threshold of $|z| = 0.86$.

Two sets of experiments were conducted on the TIMIT database,

- Training and Testing on the same telephone channel,
  - ◇ CMV–CMV
  - ◇ CPV–CPV
- Training and Testing across telephone channels,
  - ◇ CMV–CPV
  - ◇ CPV–CMV.

In either case, the results obtained were compared to those obtained by deconvolving an actual inverse estimate of the channel. The actual inverse filter was obtained from the impulse responses of the channels available from the simulator.

In the case of training and testing on the same channel, no channel compensation is necessary and the recognition accuracy is high. However, it reveals the degradation effected by ordinary CMS from the LP cepstral coefficients. As seen from Table (5.1), the recognition accuracy of identifying the speaker degrades when long-term cepstral mean is eliminated from the cepstral vectors. Hence it is evident that long-term mean removal eliminates important speaker information from the short-time spectra of every speech frame. It is shown that a performance improvement is obtained by using the long-term cepstral mean computed using pole-filtered cepstral coefficients. It is obvious that the improvement is due to the recovery of speaker information by compensating a better inverse channel estimate implicitly realized by using a refined cepstral mean estimate.

Performance of channel compensation by using pole–filtered cepstral coefficients was observed under a cross channel scenario, wherein the training data was degraded by one of the telephone channels and the testing data was degraded by the other. The CMV channel and the CPV channels were used. The performance improvement is shown in Table (5.2).

Working with known convolutional distortions allows one to decouple the cepstral mean of the training or the testing utterances into its channel component and speech component. If $\mathbf{h}_{ch}$ is the cepstrum of the known convolutional distortion and $\mathbf{s}_S$ is the cepstrum of the clean speech prior to convolution, then the channel cepstrum $\mathbf{c}_S$ for the utterance is,

$$\mathbf{c}_S \approx \mathbf{s}_S + \mathbf{h}_{ch}. \tag{5.1}$$

Method	Training	Testing	Accuracy(%)
LPCC-MR	CMV	CPV	59.4
PFCC-MR	CMV	CPV	64.7
LPCC-IC	CMV	CPV	85.3
LPCC-MR	CPV	CMV	56.8
PFCC-MR	CPV	CPV	62.6
LPCC-IC	CPV	CMV	86.8

Table 5.2: Cross channel experiment on 38 speaker subset of TIMIT using a pole-bandwidth threshold of $|z| = 0.86$.

Hence for every training or testing utterance, if the cepstral mean due to clean speech were known, one can estimate the ideal channel estimate, $\hat{\mathbf{h}}_{ch}$ as,

$$\hat{\mathbf{h}}_{ch} \approx \mathbf{c}_s - \mathbf{s}_S. \tag{5.2}$$

The relative error between the channel cepstrum $\mathbf{c}_S$ and the channel cepstrum obtained by decoupling and eliminating the clean speech ceptrum is given by,

$$Rel.Error(ordinaryCMS) = \frac{\|\mathbf{c}_S - \hat{\mathbf{h}}_{ch}\|}{\|\hat{\mathbf{h}}_{ch}\|}. \tag{5.3}$$

In the case of pole-filtered cepstral mean $\mathbf{c}_S^{pf}$, the error is given by,

$$Rel.Error(PFCMS) = \frac{\|\mathbf{c}_S^{pf} - \hat{\mathbf{h}}_{ch}\|}{\|\hat{\mathbf{h}}_{ch}\|}. \tag{5.4}$$

The relative errors have been plotted in Figure (5.1) for training utterances for ten speakers from the training set chosen for the simulated channel experiment. One can observe that the relative error due to the pole-filtered channel estimate is smaller than ordinary channel estimate.

5.1.2 Realistic channel experiments

To examine the improvement in recognition accuracy as a function of the population size and testing durations, experiments were conducted on the NTIMIT database. The experiments were made on varying population sizes of 100, 200, 326 male speakers from "train" portion of NTIMIT. The training sentences were similar to the TIMIT experiments, where all the SX sentences were concantenated. The testing for each of the population size was carried in two sets,

- Testing on individual SA and SI sentences.
- Testing after concatenating the SA and SI sentences.

The improvement in recognition accuracy when testing on individual SA and SI sentences is given in Table 5.3.

Table (5.4) shows the recognition accuracy testing on concatenated SA and SI sentences.

One can observe that the improvement in recognition accuracy using pole-filtered cepstral mean subtraction is consistent with increase in the population size and the duration of the testing utterances.

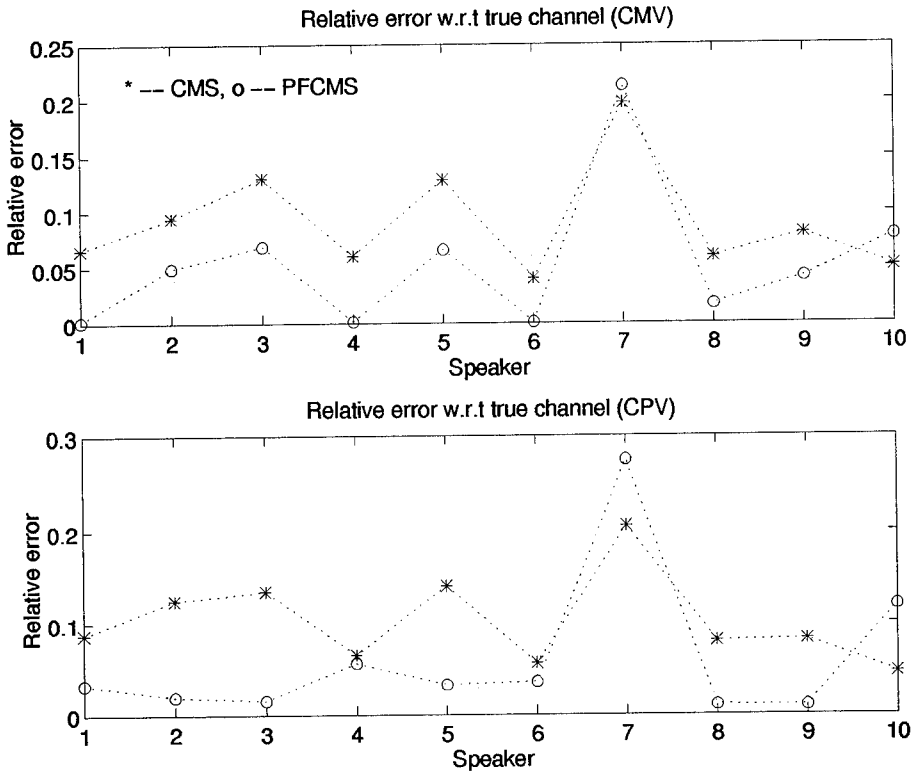


Figure 5.1: Relative error due to ordinary cepstral mean and pole-filtered cepstral mean.

Population	Clean	LPCC-MR	PFCC-MR
100	86.6	33.0	39.0
200	82.9	25.4	31.1
326	79.2	24.1	30.7

Table 5.3: NTIMIT experiments by testing on individual SA,SI sentences. Clean speech testing results obtained from the corresponding utterances in TIMIT.

5.2 Experiments on the KING database

The King Database [89] consists of speaker data recorded under realistic telephone channel environment. The database consists of 51 adult male speakers. Ten conversations of each speaker have been recorded in ten sessions over long distance telephone lines. Conversations from twenty-six subjects were recorded in San Diego, California, and the rest were recorded in Nutley, New Jersey. The first five sessions were recorded at intervals of one week and the remaining five were recorded at one month intervals. The recordings were carried out in quiet rooms. The speech was acquired via a standard carbon-button microphone after being received over a long distance transmission.

The database has been categorized into two groups consisting of five sessions each. Sessions 1-5 form one group and sessions 6-10, form the other group. Training and testing within a group has been termed as experiments *within the great divide* , and testing across the group as experiments *across the divide*, which imply considerable channel mismatch across the groups. The training for all experiments was conducted

Population	LPCC-MR	PFCC-MR
100	62.0	68.0
200	55.0	65.0
326	47.2	58.9

Table 5.4: NTIMIT experiments by testing on concantenated SA,SI sentences. Clean speech testing results obtained from the corresponding utterances in TIMIT.

on one session and the testing on the rest of the sessions. Training and testing within the great divide consisted of training on one session and testing on remaining four while experiments across the divide were conducted by training on an individual session in group 1 (sessions 1-5) and testing on all sessions in group 2 (sessions 6-10). The training durations used for these experiments is considerably smaller (the emphasis is on short duration testing and training) than those reported by Kao et.al [39] and Reynolds [58, 64], where the training data was combined from three sessions.

The experiments below have been conducted on both the San Diego portion and the Nutley portion. For the San Diego portion the experiments were conducted within and across the divide. For the Nutley portion the emphasis was placed only on the experiments across the divide. For experiments within the divide, there were 26 speakers $\times$ 4 test sessions $\times$ 5 sets = 520 test utterances, while for experiments across the divide there were 26 speakers $\times$ 5 test sessions $\times$ 5 sets = 650 test utterances in the San Diego portion. For the Nutley portion there were 25 speakers $\times$ 5 test sessions $\times$ 5 sets = 625 test utterances.

5.2.1 Preprocessing for the KING database

In the preprocessing step, a speech/silence discrimination was carried out based on thresholding of energy in the individual spoken utterance. For each utterance, the energy threshold was decided by constructing a histogram of frame energies. Only frames of energy higher than a decided threshold are considered for further processing.

Speech/non-speech discrimination based on energy, tends to remove parts of the utterance that do not correspond to voiced speech. Features extracted from these frames do not contain reliable speaker information and degrade the performance undesirably by affecting the interspeaker and intraspeaker variances. A pole-based frame selection process [31] was used to eliminate these undesirable frames for experiments across the divide. The frame selection process chooses only voiced frames of speech indicated by the number of narrow band components with bandwidths less than a pre-determined bandwidth threshold. Improvements due to the pole-based frame selection process have been reported below.

5.2.2 Similar training–testing conditions: Within the divide

For experiments within the divide, only energy thresholding was used for speech/non-speech discrimination. A pole-bandwidth threshold of $\alpha = 0.9$ was used to generate the Pole-filtered cepstral coefficients. Due to a relatively good transmission quality in sessions 1–5, no pole-based frame selection was used for experiments within the divide. Results have been compared to ordinary mean removal in Table (5.5).

5.2.3 Mismatched training–testing conditions: Across the divide

For results across the divide, pole-filtered cepstral coefficients were generated with different bandwidth thresholds. The performance across different sessions varied. Table (5.6) shows the performance improve-

Method	Identification rate	Accuracy(%)
LPCC-MR	383/520	73.6
PFCC-MR	406/520	78.1

Table 5.5: Training and testing "within the great divide" (San Diego portion).

ment when a pole-bandwidth threshold of $|z| = 0.9$ was used, for the San Diego portion.

The pole-based frame selection was then carried out in addition to energy based speech/non-speech discrimination. Further improvement in the performance was obtained using the pole selection strategy. The results were obtained by using a threshold of pole-radius $\alpha = 0.9$ for both selection of voiced frames and generating the pole-filtered cepstral coefficients. The results have been reported in Table (5.6).

Method	Identification rate	Accuracy(%)
LPCC-MR	314/650	48.2
PFCC-MR	346/650	53.2
FS + PFCC-MR	366/650	56.3

Table 5.6: Testing "across the great divide", energy based silence removal (San Diego portion).

The pole-filtered cepstral mean derived by averaging the ACW subtractive component was used to replace ordinary cepstral mean removal for exepriments across the divide. The results have been tabulated in Table (5.7).

The Nutley portion was found to exhibit poor recogntion performance due to very noisy channel conditions. However, the used of pole-filtered cepstral mean improved the recognition performance slightly as shown in Table (5.8).

5.2.4 Combined inter-intra frame approaches

It was suggested in Chapter Four that the interframe and the intraframe approaches may be considered to be complementary approaches. A two-pass approach was discussed in Chapter Four where a pole-filtered channel estimate is deconvolved from the speech utterance and then followed by conventional intraframe processing. It can be found from Tables (5.9) and (5.10) that the two approaches complement each other to yield further improvement over ordinary cepstral mean subtraction. Table (5.9) shows the performance improvement on the San Diego portion while the Table (5.10) shows the performance improvement on the Nutley portion.

Method	Identification rate	Accuracy(%)
LPCC-MR	314/650	48.2
FS + PFCC-MR(ACW)	349/650	53.7

Table 5.7: Testing "across the great divide", energy based silence removal followed by pole-based frame selection (San Diego portion).

Method	Identification rate	Accuracy(%)
LPCC-MR	184/625	29.4
FS + PFCC-MR	210/625	33.6

Table 5.8: Testing "across the great divide", energy based silence removal followed by pole-based frame selection (Nutley portion).

Method	Identification rate	Accuracy(%)
LPCC-MR	314/650	48.2
FS + PF + ACW	391/650	61.1

Table 5.9: Performance improvement "across the great divide" by combining the inter and intraframe pole filtering approaches (two-pass) (San Diego portion).

Method	Identification rate	Accuracy(%)
LPCC-MR	184/625	29.4
FS + PF + ACW	228/625	36.5

Table 5.10: Performance improvement "across the great divide" by combining the inter and intraframe pole filtering approaches (two-pass) (Nutley portion).

The computational complexity involved in a two-pass approach is more than conventional approaches and the pole filtering approach. However, a multi-pass approach is implied in reducing the effect of a convolutional distortion.

Chapter 6

Channel Identification

Besides providing an improved channel normalization scheme, the cepstral domain processing for an utterance can also be utilized for channel detection. The pole-filtered cepstral mean estimate represents a better estimate of the convolutional distortion than the ordinary cepstral mean. These channel estimates can be trained using a classifier to provide channel discrimination among significantly different channels.

The channel detection scheme has been applied here to two application scenarios:

1. To discriminate live speech from tape-recorded speech for the purpose of secure access.
2. Classify telephone handset categories for detecting channel mismatch. Experiments have been performed to classify electret phones from carbon-button phones.

6.1 Secure Access

Secure access promises to be one of the most critical issues in the deployment of speaker recognition systems in the field. During verification of a speaker's identity, a system would need a method of checking if the voice is being mimicked by an impostor intending to gain unauthorized access.

One could also expect unauthorized access attempts by using speaker information played back through a hand-held device. A common example would be when an impostor attempts to break into a system by playing back the actual speaker's voice that has been pre-recorded by using a tape recording device.

A speech-based system identification system for secure access would be helpful when discriminating between speech spoken in person (or spoken live) and speech played via a recording device. It has been observed that there are enough distortions in the played back speech that would help distinguish a spoken utterance from a recorded utterance.

Common distortions are:

- The recording device channel.
- Characteristics of the loud-speaker through which the recorded speech is being played.
- Reverberation effects from the hand-held device and the microphone used to acquire the speech.
- Nonlinearities in the sampling devices involved.

It is in general, non-trivial, to decouple these effects from the spoken utterance and identify the distortions. Either the short-time cepstral information, or the ordinary or pole-filtered cepstral mean, can be used to get an estimate of the distortions.

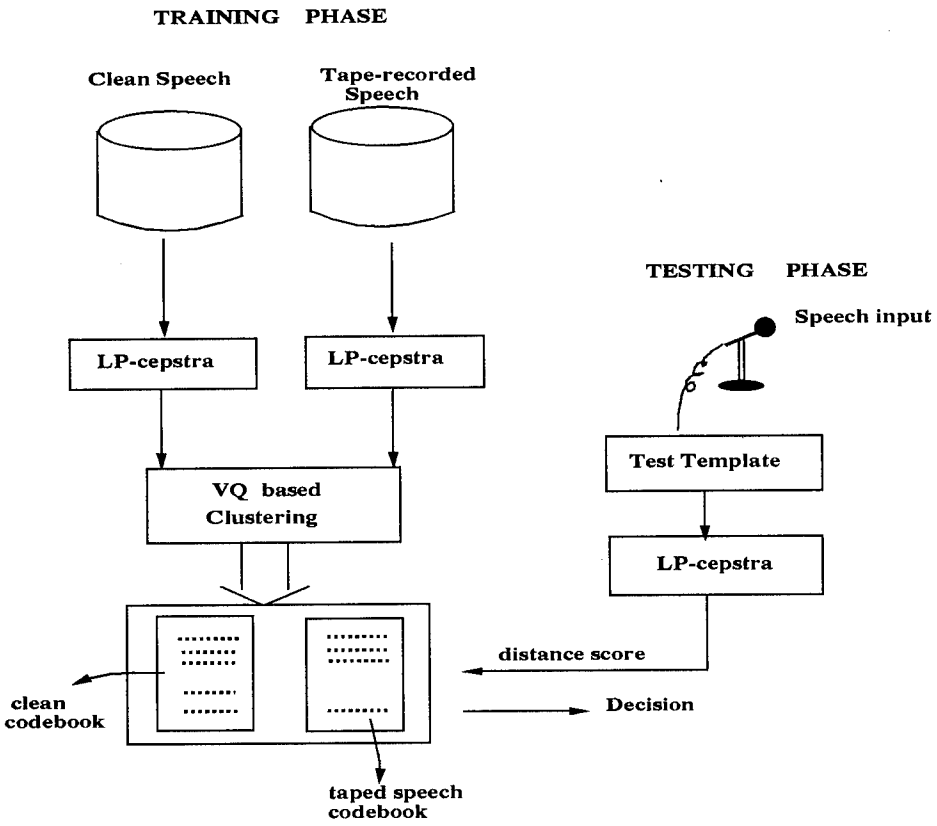


Figure 6.1: Block Diagram for speech-based system identification.

The system developed for this live/taped speech discrimination uses the the cepstral vectors extracted from the spoken utterances. The training was carried out by building a database of ten speakers. Each speaker was advised to speak five utterances of ten seconds each. While the speech was being acquired, the same utterance were also recorded into a high-quality DAT (Digital Audio Tape) recording device. The spoken utterances consisted of unconstrained text.

The database for tape-recorded speech was collected by recording the live utterances on the DAT and playing back through a loud-speaker held close to the microphone piece. LP derived cepstral vectors were collected from individual databases for clean and tape-recorded speech. A Vector Quantization (VQ) based classifier [92] was used to cluster the twelve-dimensional feature vectors into codebooks consisting of 46 centroids for each of the databases. Hence, a codebook was developed for clean speech and a similar size codebook was developed for tape-recorded speech.

The testing was carried out by evaluating a minimum L_2 distance score to the codebooks. The block diagram of the system is shown in Figure (6.1). The utterance was classified as being "live" or "tape-recorded" based on the minimum accumulated distance to each of the codebooks.

The speech–based device identification system could distinguish between tape-recorded speech and live speech with very high accuracy.

6.2 Classification of telephone handsets

It is often desirable to measure the amount of channel mismatch between training and testing utterances before performing the channel normalization step. A scheme that detects channels may be used to train

separate models for a particular speaker based on different telephone types. Two distinct telephone handset types that exhibit a significant mismatch are the Electret handset and Carbon-button handset. A reasonably accurate handset classification scheme would allow categorizing the test utterance to separate models trained on Electret and Carbon button phones.

Experiments were performed to classify Electret phones from Carbon button phones. An initial database of 12 speakers was collected. Each speaker had a total of 36 short utterances from each telephone handset category, the linear and the carbon-button types. The pole-filtered cepstral means were evaluated on parts of an utterance obtained by sliding a fixed window comprised of 80 percent of the entire utterance. This window was slid by a single sample at a time. The pole-bandwidth threshold was fixed to a value corresponding to a frequency of 300 Hz, corresponding to a $|z| = 0.88$ on the Z–plane. A Neural Tree Network classifier was trained on a part of the speaker database and testing was carried out using a set of speakers that were not present in the training set. The classification accuracy is given in Table (6.1).

Telephone type	Accuracy(%)
Linear Phone	76.0
Carbon Button	73.0

Table 6.1: Experiments on Classification of two phones.

Further experiments were conducted on a database with 35 speakers whose utterances were collected from a variety of linear phones. Carbon-button data was collected on a set of 18 speakers. The recognition accuracy is reported in Table (6.2).

Telephone type	Accuracy(%)
Linear Phone	86.3
Carbon Button	74.0

Table 6.2: Experiments on Classification of two phones categories.

Thus, an ensemble of cepstral vectors of an utterance can be utlized to get an estimate of the convolutional distortion. The method was shown to be particularly helpful for channel detection by eliminating false rejects due to channel differences for speaker verification [52].

Chapter 7

Conclusion and Future Work

7.1 Conclusion

A new philosophy for extracting robust features in speech systems, called Pole filtering, was introduced in this report. The solution of the LP difference equation of speech was interpreted in terms of poles that correspond to the eigenmodes modeling a speech segment. Consequently it was shown that the cepstral transformation is simply a homogeneous solution of the linear system implicit in the difference equation of speech. That is, the cepstral sequence can be derived as a summation of the eigenmodes of a system whose eigenvalues are the poles of the all pole filter.

The pole-filtering methodology was developed by investigating speech degraded by simulated transmission channels. Based on a study of the effects of channel variations on the poles of speech, algorithms were proposed to improve the channel estimates using an interframe processing approach. The algorithms for channel normalization were developed using the cepstral transformation of the poles of speech. A previously proposed approach by Assaleh and Mammone [31] was unified under a pole filtering methodology.

By comparing the conventional methods for channel normalization in the cepstral domain, it was emphasized that all conventional methods share a common basis for channel normalization. The basis is in effectively removing the long-term cepstral mean. It was shown that the CMS itself has a drawback, in that it corresponds to eliminating an improper cepstral estimate of the convolutional distortion. A methodology was developed wherein the deconvolution term in the cepstral domain based on ordinary long-term averaging of cepstral coefficients was refined using Pole filtering. The estimates of the channel were improved by introducing the concept of pole-filtered cepstral coefficients. Utilizing a pole-filtered cepstral mean estimate for CMS the recognition accuracy in speaker identification was improved due to a more accurate channel normalization.

The improved cepstral channel estimates obtained by pole-filtering were also used for the purpose of channel detection in Chapter Six. Channel detection experiments were performed for the purpose of

- Secure access in Speaker identification to disriminate live speech from tape-recorded speech.
- Detection of telephone handset type.

In addition to refining the cepstral mean estimate, a two-pass approach was introduced. The two-pass approach involved deconvolving the channel estimate obtained using pole-filtered cepstral mean of the speech utterance on a first pass. A conventional intraframe processing technique was then employed on the second pass. It was found that the interframe and intraframe methods complement each other and were used to further improve the recognition accuracy for speaker identification.

7.2 Future Work

The subsequent sections propose several avenues for future work in improving channel normalization utilizing the concepts investigated in pole-filtering.

Improving conventional channel normalization

Removal of the cepstral bias is the underlying basis for most cepstral channel normalization techniques. All conventional algorithms need to be re-investigated with the proposed pole-filter cepstral mean estimate to remove any bias for channel normalization instead of using the ordinary cepstral bias. The RASTA filter which was reviewed section [3.2.2], can hence be altered to modify the high pass portion that eliminates the dc component in the log spectra. An appropriate pole filtering step succeeding a modified RASTA filtering could combine advantages of RASTA and pole-filtering.

Modifying FFT Cepstral mean estimate

In the case of evaluating the LP-derived cepstral mean, channel estimates were improved by selectively modifying the all-pole parameters of the LP filter. Similar pole-filtering ideas can be applied to FFT cepstrum by applying smoothing strategies to the log-spectral magnitudes of the frames of speech to de-emphasize the speech information. One possible approach is to find a least squares polynomial fit to the magnitude response of the speech frame. After smoothing the magnitude response, a cepstral transformation can be performed. The cepstral mean of a modified cepstral transformation may represent a better channel estimate for CMS.

Channel normalization for speech recognition

Speech recognition systems often resort to sentence based or word based cepstral mean removal. Short utterance duration also affects channel compensation when using ordinary cepstral bias removal. Moreover, the presence of speech information in the cepstral mean, affects parts of the word or utterance.

By studying the spectral contents of individual phonemes for high quality (clean) speech, one can store phoneme dependent estimates of the cepstral mean, $s_S^{phoneme}$. An estimate of the underlying channel distortion can then be obtained by computing,

$$h_S^{phoneme} = c_S^{phoneme} - s_S^{phoneme}, \tag{7.1}$$

where, $c_S^{phoneme}$ corresponds to the cepstral mean from training, and $h_S^{phoneme}$ is the estimate of the distortion. During testing (or decoding), the channel estimate may have to be obtained in a maximum likelihood framework where the likelihood of the underlying speech model and the channel model are estimated.

Adapting the pole-filtering thresholds

For most experiments, the pole-bandwidth thresholds were fixed *apriori* based on an empirical estimate of the pole-bandwidths obtained by performing an all-pole LP fit to the impulse responses of the simulated distortions. The thresholds however, which control the level of smoothing in the cepstral mean, may be adaptively modified on a per utterance basis. This is done by observing the spectral content that may correspond to the speech information in the cepstral mean or the gross spectral distribution of the utterance prior to convolution.

Chapter 8

Minimum phase property of the channel compensation filter

The spectral distribution of the cepstral mean $\mathbf{c}_S$ for a sentence, S, is obtained by recursively converting the cepstral coefficients to predictor coefficients in equation [4.20]. Thus, an all-pole approximation to the channel given by, $N_{ch}(z) = \frac{1}{N_{ccf}(z)}$ is implied.

The filter, $N_{ccf}(z)$, hence corresponds to a deconvolution filter. The effect of cepstral mean subtraction is effectively a deconvolution of the channel in the time domain. It is proven below that the deconvolution filter $N_{ccf}(z)$ is an FIR filter that may be minimum phase and stable.[1] In other words, the truncated cepstral mean sequence obtained by a superposition of cepstral sequences of individual speech frames of an utterance is causal and the corresponding filter (CCF) is minimum phase.

The property is derived by investigating the all-pole filter polynomials for each frame of speech, given by $A_1(z), A_2(z), \cdots, A_M(z)$, for M frames of an utterance. These filters are minimum-phase, when derived using autocorrelation analysis, and correspond to poles inside the unit circle. Thus,

$$A_1(z) = \frac{1}{\prod_{k=1}^{P}(1 - z_{1k}z^{-1})}, \tag{8.1}$$

$$A_2(z) = \frac{1}{\prod_{k=1}^{P}(1 - z_{2k}z^{-1})}, \tag{8.2}$$

$$\vdots \tag{8.3}$$

$$A_M(z) = \frac{1}{\prod_{k=1}^{P}(1 - z_{Mk}z^{-1}).} \tag{8.4}$$

The corresponding cepstral transformation may be obtained using the root power sum formulation, given by,

$$c_1(n) = \frac{1}{n}\sum_{k=1}^{P} z_{1k}^n \longleftarrow A_1(z) \tag{8.5}$$

$$c_2(n) = \frac{1}{n}\sum_{k=1}^{P} z_{2k}^n \longleftarrow A_2(z) \tag{8.6}$$

$$\vdots \tag{8.7}$$

[1] I thank Ravi Ramchandran for showing me this proof.

$$c_M(n) = \frac{1}{n}\sum_{k=1}^{P} z_{Mk}^{n} \longleftarrow A_M(z), \tag{8.8}$$

for the n^{th} cepstral coefficient.

The corresponding summation of the cepstral sequences for the M frames in the utterance is given by,

$$\sum_{i=1}^{M} c_i(n) \longleftarrow N_{ch}(z) = \frac{1}{\prod_{k=1}^{P}(1-p_{1k}z^{-1})\prod_{k=1}^{P}(1-p_{2k}z^{-1})\cdots\prod_{k=1}^{P}(1-p_{Mk}z^{-1})}, \tag{8.9}$$

and the cepstral mean for the M frames corresponds to the geometric mean of the product of the individual all-pole filter polynomials in the frequency domain. Hence,

$$\frac{1}{M}\sum_{i=1}^{M} c_i(n) \leftrightarrow \langle N_{ch}(z).\rangle^{\frac{1}{M}} \tag{8.10}$$

One can observe from equation [8.9], $N_{ch}(z)$ is a product of minimum-phase polynomials and hence is essentially minimum phase.

An all-pole channel filter that corresponds to the cepstral mean, can be given by,

$$N_{ccf}(z) = \frac{1}{\langle N_{ch}(z)\rangle^{\frac{1}{M}}}. \tag{8.11}$$

Raising both sides of equation [8.11] to the M^{th} power one gets,

$$[N_{ccf}]^M = \prod_{k=1}^{P}(1-p_{1k}z^{-1})\prod_{k=1}^{P}(1-p_{2k}z^{-1})\cdots\prod_{k=1}^{P}(1-p_{Mk}z^{-1}), \tag{8.12}$$

which is FIR and minimum-phase of order MP.

The channel compensation filter $N_{ccf}(z)$ of order P is an approximation to the channel filter order MP in equation [8.12]. Thus the CCF is derived by truncating the impulse response of $N_{ch}(z)$. This approximation may lose its minimum phase property due the truncation of the impulse response. However, CCF has been employed only to observe the spectral distribution of the cepstral mean and does not affect the pole-filtering approach.

The filter $N_{ccf}(z)$, was also used to deconvolve the channel estimate in the first pass of a two-pass approach described in Chapter Four. The minimum phase property of the filter $N_{ccf}(z)$ is not critical here, since the deconvolution filter is FIR and hence stability is not a critical issue.

Chapter 9

Zero–mean property of cepstral coefficients

Past literature has exploited the channel invariance property of cepstral coefficients under the assumption that the cepstral coefficients are typically zero–mean under "regular conditions" [11, 73]. Based on this assumption, the conventional CMN technique has widely been claimed to provide channel normalization.

For most practical training and testing durations, the zero-mean property of cepstral coefficients is not true. For utterance durations on the order of tens of seconds, the cepstral mean of the speech utterance tends to represent a gross spectral distribution of the spoken material. When CMN is carried out, this component is eliminated along with the additive cepstral component due to the channel.

The cepstral mean of clean speech from the TIMIT [90] database is shown in the quefrency domain in Figure (9.1) and the spectral domain in Figure (9.2(a)). The cepstral mean was obtained from 10 seconds of voiced speech obtained from a section of the database. The same utterance was then degraded by a CMV channel and the cepstral mean was observed in the quefrency domain and the spectral domain. The Figures (9.1) and (9.2(b)) illustrate the changes in the cepstral and spectral domain.

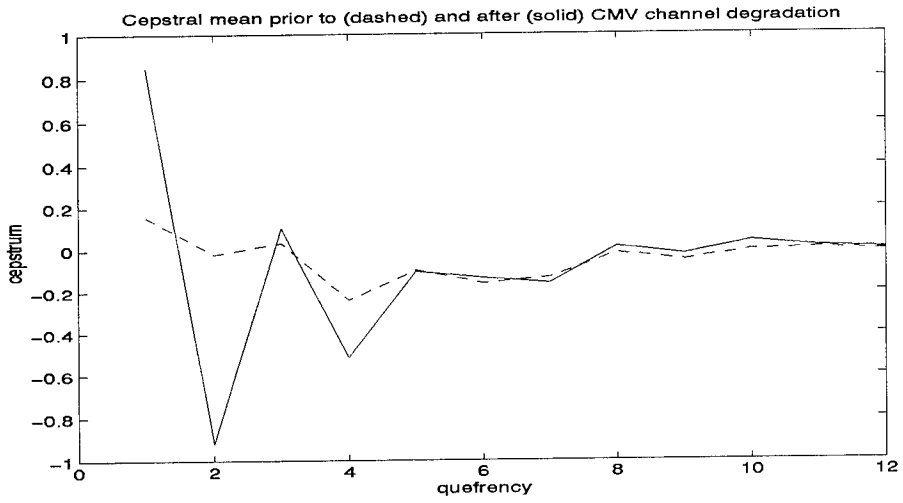


Figure 9.1: Comparison of cepstral mean of a speech utterance prior to and after channel degradation.

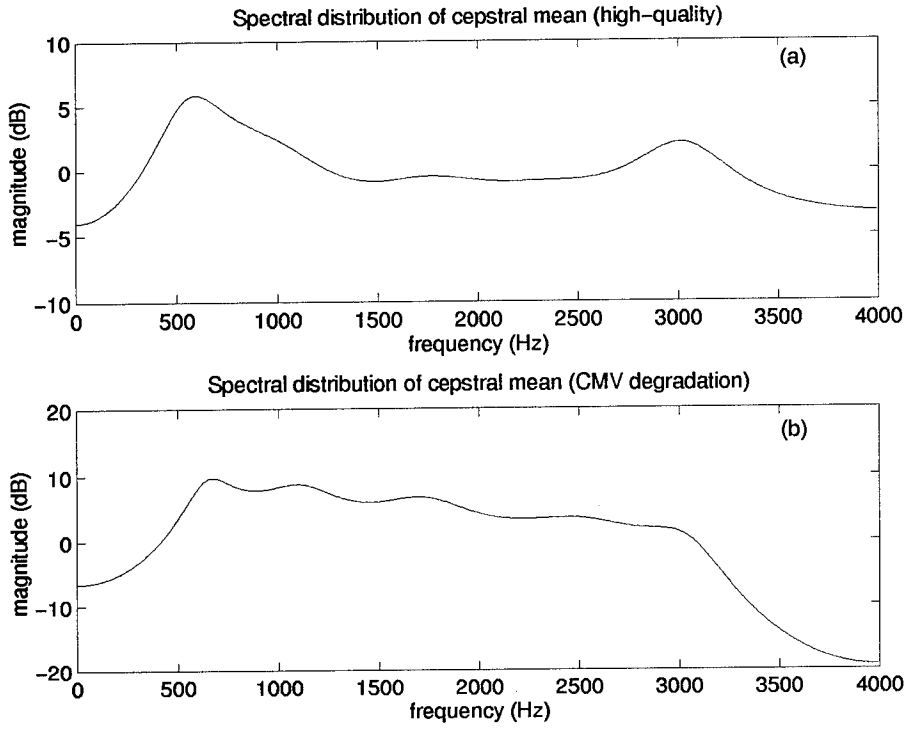


Figure 9.2: Comparison of spectra of the cepstral mean of a speech utterance prior to and after channel degradation.

Bibliography

Reviews and Tutorials

[1] B. Atal. Automatic recognition of speaker from their voices. *Proceedings of the IEEE*, 64:460–475, April 1976.

[2] A. Rosenberg. Automatic speaker recognition: A review. *Proceedings of the IEEE*, 64:475–487, April 1976.

[3] G. Doddington. Speaker recognition - Identifying people by their voices. *Proceedings of the IEEE*, 73:1651–1664, 1985.

[4] D. O'Shaughnessy. Speaker Recognition. *IEEE ASSP Magazine*, pp. 4, October 1986.

[5] J. Naik. Speaker Verification: A Tutorial. *IEEE Communications Magazine*, pp. 42, January 1990.

[6] R. Peacocke, D. Graf. An introduction to Speech and Speaker Recognition. *Computer*, pp. 26, August 1990.

[7] J. Picone. Signal Modeling techniques in Speech Recognition. *Proceedings of the IEEE*, 81:1215, September 1993.

[8] L. Rabiner. A Tutorial on Hidden Markov Models and Selected applications in Speech Recognition. *Proceedings of the IEEE* , 77:257-286, February 1989.

[9] L. Rabiner. Applications of Voice Processing to telecommunications. *Proceedings of the IEEE*, 82, pp. 199, February 1994.

[10] J. Makhoul. Linear Prediction: A Tutorial Review. *Proceedings of the IEEE*, 63:561-580, 1975.

Books

[11] L. Rabiner and B. H. Juang. *Fundamentals of Speech Recognition.* Prentice Hall, Englewood Cliffs, NJ, 1993.

[12] B. Atal. *Linear Prediction of Speech.* Computer Speech Processing, Ed. F. Fallside and W. Woods, Prentice Hall, Englewood Cliffs, NJ, 1993.

[13] J. Markel, A. Gray Jr. *Linear Prediction of Speech.* Communications and Cybernetics 12, 1982, Springer-Verlag, Berlin.

[14] L. Rabiner, R. Schafer. *Digital processing of Speech Signals.* Prentice-Hall, New Jersey, 1978.

[15] R. Duda, P. Hart. *Pattern Classification and Scene analysis.* Wiley, New York, 1973.

[16] O. Tosi. *Voice Identification: Theory and Legal Applications.* Univ. Park Press: Baltimore, MD, 1979.

[17] J.L. Flanagan. *Speech Analysis, Synthesis, and Perception.* Springer-Verlag, 1983.

[18] A. Oppenheim, R. Schafer. *Digital Signal Processing.* Chapter 10, Prentice Hall, New Jersey, 1975.

[19] R. Gabel, R. Roberts. *Signals and Linear Systems.* John Wiley and Sons Inc., 3rd Ed., 1973, New York.

[20] S. Haykin. *Adaptive filter theory.* Prentice Hall, 2nd Ed., 1991, New Jersey.

[21] D. E. Rumelhart and J. L. McClelland. *Parallel and Distributed Processing Vol. 1.* MIT press, Cambridge, MA, 1986.

[22] S. L. Marple, Jr. *Digital Spectral Analysis with Applications.* Prentice-Hall, New Jersey, 1987.

[23] W. Press, S. Teukolsky, W. Vetterling, and B. Flannery. *Numerical Recipes in C, The Art of Scientific Computing*, 2nd edition, Cambridge University Press, 1992.

Speaker Recognition

[24] K. Farrell, R. Mammone, K. Assaleh. Speaker Recognition Using Neural Networks and Conventional Classifiers. *IEEE Trans. Speech and Audio Proc.*, Vol. 2-1, part-2, pp. 2, January 1994.

[25] S. Das, W. Mohan. A Scheme for Speech Processing in Automatic Speaker Verification. *IEEE Trans. Audio Electroacoustics*, Vol. AU-19, pp. 32-43, March 1971.

[26] J. Naik, L. Netsch, and G. Doddington. Speaker verification over long distance telephone lines. *Proc. ICASSP*, pp. 524-527, 1989.

[27] J. Naik. Speaker verification over the telephone network: databases, algorithms and assessment. *ESCA workshop on Automatic Speaker recognition, identification and verification*, pp. 31, April 1994.

[28] H. Gish, K. Karnofsky, M. Krasner, S. Roucos, R. Schwartz, and J. Wolf. Investigation of speaker identification over telephone channels. *Proc. ICASSP*, pp. 379-382, 1985.

[29] B. Atal. Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification. *Journal of the Acoustical Society of America*, 55:1304–1312, June 1974.

[30] K. Assaleh, R. Mammone. Robust features for Speaker recognition. *Proc. ICASSP*, April 1994.

[31] K. Assaleh, R. Mammone. New LP-derived features for Speaker identification. *IEEE Trans,. on Speech and Audio*, Vol. 2-4, pp. 630, October 1994.

[32] S. Furui. Cepstral analysis technique for automatic speaker verification. *IEEE ASSP*, 29:254–272, April 1981.

[33] S. Furui. An analysis of long-term variations of feature parameters of speech and its application to talker recognition. *Electron. Comm.*, Vol. 57-A, pp. 34-42, 1974.

[34] F.K. Soong, A.E. Rosenberg, L.R. Rabiner, and B.H. Juang. A vector quantization approach to speaker recognition. *Proc. ICASSP*, pages 387–390, 1985.

[35] F. Soong, A. Rosenberg. On the Use of Instantaneous and Transitional Spectral Information in Speaker Recognition. *IEEE ASSP*, 36-6, June 1988.

[36] G. Velius. Variants of cepstrum based speaker identity verification. In *Proc. ICASSP*, pages 583–586, 1988.

[37] D. Reynolds. A Gaussian Mixture Modeling Approach to Text-Independent Speaker Identification. *PhD thesis*, Georgia Institute of Technology, February 1993.

[38] K. Li and E. Wrench. An approach to text-independent speaker recognition with short utterances. *Proc. ICASSP*, pages 555–558, 1983.

[39] Y. Kao, J. Baras, P. Rajasekaran. Robustness Study of Free-Text Speaker Identification and Verification. *Proc. ICASSP*, Vol. 2, pp. 379, 1993, Minneapolis.

[40] A. Higgins, L. Bahler, J. Porter. Voice Identification using nearest-neighbor distance measure. *Proc. ICASSP* , Vol 2, pp. 375, 1993, Minneapolis.

[41] H. Gish, M. Schmidt. Text-Independent Speaker Identification. *IEEE Signal Processing Magazine*, pp. 18, October 1994.

[42] M. Hunt, J. Yates, J. Bridle. Automatic Speaker recognition for user over communication channels. *Proc. ICASSP*, pp. 764-767, May 1977.

[43] M. Hunt. Further Experiments in Text-Independent Speaker Recognition over Communications channels. *Proc. ICASSP*, pp. 563-566, Boston 1983.

[44] J. Paul, A. Rabinowitz, J. Riganati, J. Richardson. Development for Analytical methods for a semi-automatic speaker identification system. *Carnahan Conf. on Crime Countermeasures*, pp. 52-64, May 1975.

[45] D. Naik, K. Assaleh and R. Mammone. Robust speaker identification using pole filtering. *ESCA workshop on Automatic Speaker recognition, identification and verification*, Martigny, Switzerland, April 1994.

[46] D. Naik, R. Mammone. Channel normalization using Pole-filtered Cepstral Mean Subtraction. *Proc. SPIE*, Vol. 2277, July 1994.

[47] D. Naik. Pole-filtered Cepstral Mean Subtraction. *Proc. ICASSP*, Accepted, May 1995.

[48] K. Farrell, S. Kosonocky, R. Mammone. Neural Tree Network/Vector Quantization Probability Estimators for Speaker Recognition. *IEEE Workshop on Neural Networks for Signal Processing*, Ermioni, Greece, September 6-8, 1994.

[49] K. Farrell, R. Mammone. Hybrid Vector Quantization/Neural Tree Network classifiers for speaker recognition, *IEEE Workshop on Neural Networks for Signal Processing*, Ermioni, Greece, September 6-8, 1994.

[50] M. Sambur. Selection of Acoustic features for Speaker Identification. *IEEE ASSP*, Vol. 23-2, April 1975.

[51] H. Wang, M. Chen, T Yang. A novel approach to speaker identification over telephone networks. *Proc. ICASSP*, 1992.

[52] K. Assaleh, K. Farrell, M. Zilovic, M. Sharma, D. Naik and R.Mammone. Text dependent speaker verification using data fusion and channel detection. *Proc. SPIE*, Vol. 2277, San Deigo, July 1994.

[53] H. Hollien, W. Majewski. Speaker Identification by long-term spectra under normal and distorted speech conditions. *Journal Acoustical Society of America*, Vol. 62, pp. 975-980, 1977.

[54] P. Bricker, R. Gnandesikan, M. Mathews, S. Puranzsky, P. Tukey, W. Wachter, J. Warner. Statistical Techniques for Talker Identification. *Bell System Tech. Journal*, Vol. 50-4, April 1971.

[55] J. Wolf. Efficient Acoustic Parameters for Speaker Recognition. *Journal of Acoustical Society of America*, Vol. 51-6(2), pp. 2044-2056, 1972.

[56] T. Matsui, S. Furui. Comparison of text-independent speaker recognition methods using VQ-distortion and discrete/continuous HMM's. *IEEE Trans. on Speech and Audio proc.*, Vol. 2-3, 1994.

[57] R. Rose, E. Hofstetter, D. Reybolds. Integrated Models of Signal and background with application to Speaker Identification in Noise. *IEEE Trans. on Speech and Audio proc.*, Vol. 2-2, 1994.

[58] D. Reynolds. Large population Speaker recognition using Wideband and Telephone Speech. *Proc. SPIE*, Vol. 2277, July 1994.

[59] R. Rose, D. Reynolds. Text independent speaker identification using automatic acoustic segmentation. *Proc. ICASSP*, pp. 293, 1990.

[60] M. Savic, S. Gupta. Variable parameter speaker verification system based on hidden markov modeling. *Proc. ICASSP*, pp. 281, 1990.

[61] J. Oglesby, J. Mason. Optimization of neural models for speaker identification. *Proc. ICASSP*, pp. 261-264, 1990.

[62] Y. Bennani, P. Gallinari. On the use of TDNN-extracted feature information in talker identification. *Proc. ICASSP*, pp. 385-388, 1991.

[63] J. Oglesby, J. Mason. Radial basis function networks for speaker recognition. *Proc. ICASSP*, pp. 393-396, 1991.

[64] D. Reynolds. Experimental Evaluation of features for robust speaker identification. *IEEE Trans,. on Speech and Audio processing*, Vol. 2-4, 1994.

Speech Recognition

[65] L. Rabiner, S. Levinson, A. Rosenberg, J. Wilpon. Speaker Independent Recognition of Isolated Words Using Clustering Techniques. *IEEE ASSP*, 27:588-595, December 1979.

[66] K. Lee, H. Won, D. Reddy. An Overview of the SPHINX Speech Recognition System. *IEEE ASSP*, 38:600-610 , 1990.

[67] Y. Tohkura. A weighted cepstral measure for speech recognition. *IEEE ASSP*, 35:947-954, October 1987.

[68] H. Hermansky, N. Morgan, A. Bayya, and P. Kohn. RASTA-PLP Speech Analysis Technique. *Proc. ICASSP*, I:121–124, San Francisco, 1992.

[69] S. B. Davis and P. Mermelstien. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE ASSP*, 28:357-366, 1980.

[70] E. Zwicker, E. Terhardt. Analytical expressions for critical-band rate and critical bandwidth as a function of frequency. *Journal of Acoustical Society of America* Vol. 68, no. 5, pp. 1523-1525, December 1980.

[71] B. Hanson, T. Applebaum. Subband and Cepstral domain filtering for recognition of lombard and channel-distorted speech. *Proc. ICASSP*, II-79, 1993.

[72] K. K. Paliwal. On the performance of the quefrency weighted cepstral coefficients in vowel recognition. *Speech Communications*, 1:151-154, 1982.

[73] B. Juang, L. Rabiner, J. Wilpon On the use of Bandpass Liftering in Speech recognition. *IEEE ASSP*, 35:947-954, July 1987.

[74] M. Rahim, B. Juang. Signal Bias Removal for robust telephone based speech recognition in adverse environments. *Proc. ICASSP*, June 1994, Australia.

[75] F. Itakura. Minimum Prediction Residual Principle applied to Speech recognition. *IEEE ASSP*, Vol. 23–1, February 1975.

[76] F. Liu, A. Acero, R. Stern. Efficient Joint compensation of speech for the effects of additive noise and linear filtering. *Proc. ICASSP*, I-257, 1992.

[77] A. Sankar, C. Lee. Stochastic Matching for Robust Speech Recognition. *IEEE signal processing letters*, Vol. 1-8, August 1994.

[78] A. Waibel. Modular Construction of Time Delay Neural Networks for Speech Recognition. *Neural Computation*, 1, March 1989.

[79] H. Hermansky, N. Morgan. RASTA processing of speech. *IEEE Trans. on Speech and Audio processing*, Vol. 2-4, 1994.

[80] L. Neumeyer, V. Digalakis, M. Weintraub. Training Issues and channel equalization techniques for construction of telephone acoustic models using a high-quality corpus. *IEEE Trans. on Speech and Audio processing*, Vol. 2-4, 1994.

Miscellaneous

[81] B. Atal. Speech Analysis/Synthesis by Linear Prediction of the Speech wave. *Journal of Acoustical Society of America*, 50, pp. 637-655, 1971.

[82] P. Crozier, B. Cheetham. C. Holt, E. Munday. Speech Enhancement employing Spectral Subtraction and Linear Predictive Analysis. *Electronics Letters*, Vol. 29-12, June 1993.

[83] S. Boll Suppression of Acoustic Noise in Speech Using Spectral Subtraction. *IEEE ASSP*, Vol. 27-2, April 1979.

[84] Y. Linde, A. Buzo, R. Gray. An algorithm for vector quantizer design. In IEEE Trans. Comm., Vol. COM-28, no. 1, pp. 84-95, January 1980.

[85] R. Gray. Vector Quantization. *IEEE ASSP Magazine*, Vol. 1, pp. 4-29, 1984.

[86] A. Sankar and R.J. Mammone. Growing and pruning neural tree networks. *IEEE Trans. on Computers*, C–42:21–229, 1992.

[87] M. R. Schroeder. Direct (nonrecursive) relations between cepstrum and and predictor c oefficients. *IEEE ASSP*, 29:297–301, April 1981.

[88] J. Kupin, A wireline Simulator [Software], CCR-P, April 1993.

[89] ITT Aerospace/Communications Division, King database.

[90] DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus. National Institute of Standards and Technology (NIST).

[91] C. Jankowski, A. Kalyanswamy, S. Basson and J. Spitz. NTIMIT: a phonetically balanced, continuous speech telephone bandwidth speech database. *Proc. ICASSP*, pp. 109, April 1990.

[92] T. Kohonen et.al. Learning Vector Quantization (LVQ–PAK) package, version 2.0. *Helsinki University of Technology*, 1991, Finland.

[93] B. Atal, J. Remde. A new model of LPC excitation for producing natural sounding speech at low bit rates. *Proc. ICASSP*, pp. 614–617, France, May 1982.

[94] M. Sharma, R. Mammone. Neural Tree Network (NTN) for speech segmentation into sub-word acoustic units. *SPIE Conference on Neural networks and Artificial Intelligence*, October 1993, Innsbruck, Austria.

[95] A. Dempster, N. Laird, D. Rubin. Maximum Likelihood from incomplete data via the EM algorithm. *Journal of Royal Statistical Society*, Vol. 39-1, pp. 1-38, 1977.

[96] H. Kunzel Current approaches to forensic speaker recognition. *ESCA workshop on Automatic Speaker recognition, identification and verification*, pp. 135, April 1994.

[97] J. Garcia, J. Rodriguez Robust speech modeling for speaker identification in forensic acoustics. *ESCA workshop on Automatic Speaker recognition, identification and verification*, pp. 217, April 1994.

[98] D. Broomhead, D. Lowe. Multivairable functional interpolation and adaptive networks. *Complex systems*, Vol. 3, pp. 269-303, 1988.

[99] D. Childers, D. Skinner, R. Kemerait. The Cepstrum: A Guide to Processing. *Proc. of the IEEE*, Vol. 65-10, October 1977, pp. 1428.

[100] T. Stockham Jr., T. Cannon, R. Ingebresten. Blind Deconvolution through Digital Signal Processing. *Proc. of the IEEE*, Vol. 63-4, pp.678, April 1975.

[101] T. Stockham, Jr. Restoration of old acoustic recordings by means of digital signal processing. Preprint, 41st Convention, *Audio Engineering Society*, NY, October 1971.

[102] S. Kuo and R. Mammone. A iterative projection technique for blind image restoration. *Journal of Visual Comm. and Image Rep.*, Vol. 4, no.1, pp. 92-101, 1993.

[103] M. Cannon. Blind deconvolution of spatially invariant image blurs with phase. *IEEE ASSP*, Vol. 24, pp. 58-63, February 1976.

[104] A. Benvineste, M. Goursat, G. Ruget. Robust identification of a nonminimum phase system: blind adjustment of a linear equalizer in data communications. *IEEE Trans. Autom. Control*, Vol. AC-25, pp. 385-399, 1980.

[105] S. Bellini. Busgang techniques for blind equalization. *Globecom*, pp. 1634-1640, Houston, Texas, 1986.

[106] A.V. Oppenheim and R.W. Schafer. Homomorphic analysis of speech. *IEEE Trans. Audio and Electroacoustics*, AU-16:221–226, June 1968.

☆U.S. GOVERNMENT PRINTING OFFICE: 1996-710-126-201 47

MISSION

OF

ROME LABORATORY

Mission. The mission of Rome Laboratory is to advance the science and technologies of command, control, communications and intelligence and to transition them into systems to meet customer needs. To achieve this, Rome Lab:

a. Conducts vigorous research, development and test programs in all applicable technologies;

b. Transitions technology to current and future systems to improve operational capability, readiness, and supportability;

c. Provides a full range of technical support to Air Force Materiel Command product centers and other Air Force organizations;

d. Promotes transfer of technology to the private sector;

e. Maintains leading edge technological expertise in the areas of surveillance, communications, command and control, intelligence, reliability science, electro-magnetic technology, photonics, signal processing, and computational science.

The thrust areas of technical competence include: Surveillance, Communications, Command and Control, Intelligence, Signal Processing, Computer Science and Technology, Electromagnetic Technology, Photonics and Reliability Sciences.