FOURIER TRANSFORM PHASE-BASED FEATURES FOR SPEECH RECOGNITION

A THESIS

submitted by

RAJESH MAHANAND HEGDE

for the award of the degree

of

DOCTOR OF PHILOSOPHY

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY MADRAS.

JULY 2005

To my lovable son
Vinayaka

THESIS CERTIFICATE

This is to certify that the thesis entitled **Fourier Transform Phase-Based Features for Speech Recognition** submitted by **Rajesh Mahanand Hegde**, to the Indian Institute of Technology, Madras for the award of the degree of Doctor of Philosophy, is a bonafide record of the research work carried out by him under my supervision. The contents of this thesis, in full or in parts, have not been submitted to any other Institute or University for the award of any degree or diploma.

Date:
Chennai 600 036. (Dr. Hema A. Murthy)

ACKNOWLEDGMENTS

I have always believed that independence of thought and freedom to imagine breed new ideas. I was extremely fortunate to have a research adviser who granted me both. Dr. Hema A. Murthy my research adviser, will always be remembered by me as a person with a professional attitude to research. I will continue to relish some of the kudos received from her even when I did not deserve one, which kept me motivated through the period of my research. It would be most appropriate to say that her initial work on the modified group delay function laid the foundation for this thesis. My sincere thanks to her for the support and encouragement received over the last three years.

I wish to thank Dr. V.R.R Gadde of SRI International, CA, for his valuable suggestions and inputs to the thesis.

Prof. Timothy A. Gonsalves is an excellent teacher and the most unassuming individual I have ever come across. I was fortunate to be associated with him by working in DON lab.

I wish to express my thanks to Prof. B. Yegnanarayana and Dr. C. Chandra Sekhar for the valuable feedback received during the formal and informal interactions with them.

I am grateful to the Head of the department, Prof. S. Raman for extending me all the administrative and official support during the last three years.

I would like to place on record my gratitude to all the members of my Doctoral committee who have constructively contributed to this thesis.

I have always cherished the technical discussions and visits to the canteen with my colleague Nagarajan. I have also enjoyed discussing with Surya and will always remember the outings with Surya, Kumaraswamy, and CKM. My visit to Singapore and S. Korea during the course of my research work, with DON Labber Sridhar will be fondly remembered by me.

I wish to thank Deivapalan, Kasirajan, and Rao of DON lab for all the help received during my seminars and with some sticky Linux stuff. I wish to thank Kareem for his insightful inputs. I am thankful for the support received from the office staff of the computer science department.

I would like to thank my parents and wife Sucheta for bearing with my antics and impositions during the period of my research. The joy of my life, Vinayaka, was a calming influence on me during phases of agitation in my mind. Finally I would like to thank God the almighty who has blessed me with the power to think and to articulate my thoughts.

Rajesh M. Hegde

ABSTRACT

Keywords: *Feature extraction, Phase spectrum, Group delay function, Feature combination, Multi-stream, Multi-band, Gaussian mixture models, Hidden Markov Models.*

Spectral representation of speech is complete when both the Fourier transform magnitude and phase spectrum are specified. In conventional speech recognition systems, features are generally derived from the short time magnitude spectrum. Although the importance of Fourier transform phase in speech perception has been realized [1], few attempts have been made to extract features from it. This is primarily because the resonances of the speech signal which manifest as transitions in the phase spectrum are completely masked by the wrapping of the phase spectrum. Hence an alternative to process the Fourier transform phase for extracting speech features, is to process the group delay function which can be directly computed from the speech signal [2]. The focus of this thesis is to extract single and multi-stream speech features from the modified group delay function, which is based on the Fourier transform phase, and to use them for various speech recognition applications.

A new method of feature extraction which is based on the negative derivative of the Fourier transform phase spectrum, also called the group delay function, is described in this thesis. The group delay function can be used to accurately represent signal information as long as the roots of the z-transform of the signal are not too close to the unit circle in the z-plane. Further the presence of pitch zeros also makes the group delay function spiky. The computation of the group delay function is modified to overcome these effects. Cepstral features are extracted from the modified group delay function using the discrete cosine transform (DCT) and are called the modified group delay feature (MODGDF). Velocity and acceleration parameters for the modified group delay feature are computed in the cepstral domain, in a conventional manner and appended to the MODGDF.

The MODGDF is then evaluated using several feature evaluation criteria like decorrelation, robustness to convolutional and white noise, task independence, feature selection and class discrimination. The feature evaluation results are compared with that of the Mel frequency cepstral co-efficients, which are perhaps the most commonly used speech features. The results of performance evaluation of the MODGDF for three speech recognition tasks namely automatic speaker identification (on the TIMIT [3] and NTIMIT [4] databases), automatic language identification (on the *OGI_MLTS* database), and continuous speech recognition (on the DBIL database [5]) are presented. These results are also compared with the spectral and root compressed forms of the MFCC and linear frequency cepstral co-efficients (LFCC).

The performance of speech recognition systems can be improved by application of the multi-stream and multi-band paradigm (MS-MB) to ASR [6,7]. Conventional multi-stream and multi-band (MS-MB) approaches combine multiple streams of features derived from narrow frequency bands over the entire frequency band using early fusion or late fusion.

In this thesis, a new approach within the multi-stream and multi-band framework is proposed. The whole frequency band is divided into multiple frequency bands. The modified group delay feature derived from the Fourier transform phase and the MFCC derived from the Fourier transform magnitude are extracted across multiple frequency bands. The two feature streams are then recombined across selective frequency bands using early fusion. This approach is called the multi-stream and selective-band (MS-SB) approach. The approach is also extended to combining features extracted from the entire frequency band with the features extracted from the MS-SB approach. This is called the multi-stream with full and selective band (MS-FA-SB) approach. Both these approaches are used in the recognition of continuous speech from the DBIL database [5]. Significant improvements in recognition performance, in the range of 13-14% are achieved over the conventional single-stream full-band (SS-FB) approach.

TABLE OF CONTENTS

Thesis certificate **i**

Acknowledgments **ii**

Abstract **iv**

List of Tables **xii**

List of Figures **xiv**

Abbreviations **xxi**

1 INTRODUCTION **1**

1.1 Motivation 1
1.2 Importance of Phase in Speech Processing 2
1.3 Objectives of the Thesis 5
1.4 Organization of the Thesis 5
1.5 Major Contributions of the Thesis 8

2 BACKGROUND TO SPEECH FEATURE EXTRACTION **10**

2.1 Feature Extraction versus Classification 10
2.2 The Speech Feature Extraction Paradigm 11
2.2.1 Spectral Shaping 11
2.2.2 Spectral Analysis 12
2.2.3 Feature Transformation 14
2.3 Review of Approaches to Speech Feature Extraction 14
2.4 Some Popular Feature Extraction Techniques 16
2.4.1 Linear Prediction Cepstral Co-efficients 17
2.4.2 Perceptual Linear Prediction Co-efficients 18

2.4.3 Linear Frequency Cepstral Co-efficients 19
2.4.4 Mel Frequency Cepstral Co-efficients 20
2.5 Summary 22

3 THEORY AND PROPERTIES OF GROUP DELAY FUNCTIONS 23
3.1 The Group Delay Function 23
3.2 Group Delay Spectrum and Magnitude Spectrum 24
3.3 Relationship between Spectral Magnitude and Phase 25
3.4 The Restriction of Minimum Phase 26
3.5 Properties of Group Delay Functions 27
3.5.1 Additive Property 27
3.5.2 High Resolution Property 28
3.6 Formant Analysis using The Group Delay Function 30
3.7 Summary 33

4 THE MODIFIED GROUP DELAY FEATURE 34
4.1 Significance of Zero Proximity to the Unit Circle 34
4.2 Significance of Pitch Periodicity Effects 36
4.3 The Modified Group Delay Function 37
4.3.1 Significance of Cepstral Smoothing 38
4.3.2 Definition of the Modified Group Delay Function 40
4.4 Parameterizing The Modified Group Delay Function 42
4.4.1 Importance of c_0 42
4.4.2 Algorithm for Computing The Modified Group Delay Cepstra . 43
4.5 Summary 44

5 FEATURE EVALUATION AND ANALYSIS 45
5.1 Decorrelation 45
5.2 Robustness 45
5.2.1 Robustness to Convolutional and White Noise 46
5.2.2 Comparison to Log and Root Compressed Cepstra 47
5.2.3 Similarity to RASTA 49
5.2.4 Significance of Cepstral Mean Subtraction 50
5.2.5 Significance of Removing Channel Effects in the Group Delay Domain 50
5.3 Analysis of Task Independence using the Sequential Forward Search . . 51
5.4 Cluster Structure and Separability Analysis in the 2-Dimensional Feature Space 53
5.4.1 Separability Analysis using Sammon Mapping 54
5.5 Separability Analysis in the High-Dimensional Feature Space 56
5.6 Summary 58

6 PERFORMANCE EVALUATION 59
6.1 Databases used in the Study 59
6.1.1 The TIMIT Database 60
6.1.2 The NTIMIT Database 60
6.1.3 The OGI_MLTS Database 60
6.1.4 The Database for Indian Languages (DBIL) 61
6.2 Computation of Various Features 61
6.2.1 Algorithm for Computing the Modified Group Delay Cepstra . . 61
6.2.2 Extraction of MFCC 62
6.2.3 Extraction of Spectral root and Energy Root Compressed MFCC 63
6.3 Estimation of Optimal Values for $lifter_w$, α, and γ 64
6.3.1 Estimation of Optimal Values for $lifter_w$ 64
6.3.2 Estimation of Optimal Values for α and γ 66
6.3.3 Estimation of Optimal Values for $lifter_w$, α, and γ using Line Search 68
6.4 Baseline System and Experimental Results for Automatic Speaker Identification 69
6.4.1 Discussion 71

6.5 Baseline System and Experimental Results for Language Identification 71
  6.5.1 Discussion . . . 71
  6.5.2 Confusability Analysis for the 11 Language Task . . . 73
6.6 Baseline System and Experimental Results for Syllable Recognition . . 75
  6.6.1 Overview of The Segment-Based Approach to Recognition of Continuous Speech . . . 75
  6.6.2 Issues in Segment-Based Recognition of Continuous Speech . . . 77
  6.6.3 Tri-Syllabic Viterbi Re-Alignment . . . 78
  6.6.4 Extension of The Tri-Syllabic Viterbi Re-alignment to 15-best Alternatives . . . 81
  6.6.5 The Baseline System . . . 82
  6.6.6 Discussion . . . 83
6.7 Significance of Feature Combinations . . . 85
  6.7.1 Feature Combination Before the Acoustic Model . . . 85
  6.7.2 Likelihood Combination After the Acoustic Model . . . 85
6.8 Reconstruction of Formants from the MODGDF, MFCC, and Joint features . . . 86
  6.8.1 Algorithm for Reconstruction of the Short-Time Power Spectra from the MFCC . . . 87
  6.8.2 Algorithm for Reconstruction of Short-Time Modified Group Delay Spectra from the MODGDF . . . 88
  6.8.3 Algorithm for Reconstruction of Short-Time Composite Spectra from Joint Features (MODGDF+MFCC) . . . 88
  6.8.4 Formant Reconstruction for a Synthetic Vowel . . . 89
  6.8.5 Formant Reconstruction for Clean and Noisy Speech Data . . . 91
6.9 Experimental Results for Combined Features . . . 93
  6.9.1 Extraction of Joint Features Before the Acoustic Model . . . 93
  6.9.2 Likelihood Combination After the Acoustic Model . . . 94
  6.9.3 Significance of Combining Other Features . . . 94

  6.9.4 Experimental Results of Combined Features for Speaker Identification . . . 95
  6.9.5 Experimental Results of Combined Features for Language Identification . . . 96
  6.9.6 Experimental Results of Combined Features for Syllable Based Continuous Speech Recognition . . . 98
6.10 Discussion . . . 98

7 THE MULTI-STREAM AND SELECTIVE-BAND PARADIGM 101

7.1 Psycho-Acoustic Motivation to Multi-stream and Multi-band Speech Processing . . . 102
7.2 Review of Multi-stream Speech Processing . . . 102
7.3 Review of Multi-band Speech Processing . . . 104
7.4 The Multi-stream and Multi-band paradigm . . . 104
  7.4.1 The Multi-stream Paradigm . . . 105
  7.4.2 The Multi-band Paradigm . . . 105
  7.4.3 Mathematical Interpretation of the Product of Errors Rule for the MS-MB Formulation . . . 106
7.5 The Multi-stream and Selective-band Paradigm . . . 107
  7.5.1 The single-stream (SS-FB) paradigm . . . 108
  7.5.2 The multi-stream and full-band (MS-FB) paradigm . . . 108
  7.5.3 The single-stream and multi-band (SS-MB) paradigm . . . 109
  7.5.4 The multi-stream and selective band (MS-SB) paradigm . . . 111
    7.5.4.1 Analytical Formulation of the MS-SB approach . . . 111
    7.5.4.2 Extension of the MS-SB approach to the multi-stream-full-and-selective band (MS-FASB) approach . . . 112
7.6 Performance Evaluation . . . 113
  7.6.1 Databases and Baseline System used in the Study . . . 114
  7.6.2 Recognition Accuracy Evaluation Methodology . . . 114

7.6.3 Experimental Results for the Single-Stream Full-Band and Single-Stream Multi-Band Approach 114
7.6.4 Experimental Results for the Multi-Stream Full-Band and Multi-Stream Selective-Band Approach 115
7.6.5 Experimental Results for the Multi-Stream Full and Selective-Band Approach 116
7.7 Discussion 117

8 SUMMARY AND CONCLUSIONS **119**
8.1 Summary 119
8.2 Key Ideas Presented in The Thesis 121
8.3 Criticisms 121
8.4 Scope for Future Work 122

Appendix A **123**

Appendix B **126**

Appendix C **129**

Appendix D **132**

References **136**

LIST OF TABLES

6.1 Series of experiments conducted on various databases with the MODGDF. 69
6.2 Best front-end for the MODGDF across all tasks and across all databases used in the study. 69
6.3 Recognition performance of various features for speaker identification. MODGDF (MGD), MFCC (MFC), LFCC (LFC), spectral root compressed MFCC (SRMFC), energy root compressed MFCC (ERMFC), and spectral root compressed LFCC (SRLFC) 70
6.4 Recognition performance of various features for language identification. MODGDF (MGD), MFCC (MFC), LFCC (LFC), spectral root compressed MFCC (SRMFC), energy root compressed MFCC (ERMFC), and spectral root compressed LFCC (SRLFC) 72
6.5 Confusion matrix for the 11 language task, E: English, Fr: French, K: Korean, M: Mandarin, Fa: Farsi, G: German, S: Spanish, H: Hindi, V: Vietnamese, T: Tamil and J: Japanese. 73
6.6 Five best alternatives of each recognized segment for the utterance */I rOju vArtallOni mukhyAmshAlu/* using the baseline system. 78
6.7 Five best alternatives of each recognized segment for the utterance */SIL dE si ya vA da kAZN gras/* using the baseline system. 82
6.8 Recognition performance of various features for syllable recognition. MODGDF (MGD), MFCC (MFC), LFCC (LFC), spectral root compressed MFCC (SRMFC), energy root compressed MFCC (ERMFC), and spectral root compressed LFCC (SRLFC) 84
6.9 Recognition performance of combined features for speaker identification. 96
6.10 Recognition performance of combined features for language identification. 97
6.11 Recognition performance of combined features for syllable recognition. . 99

7.1 Baseline syllable recognition accuracy (SRA) and the conventional HMM (HTK) based syllable recognition accuracy (HTKSRA) of the MFC: MFCC, MGD: MODGDF and joint features for the DBIL data. ($sb1$ is the lower sub-band and $sb2$ is the upper sub-band). 115
7.2 Baseline syllable recognition accuracy (SRA) and the conventional HMM (HTK) based syllable recognition accuracy (HTKSRA) of the joint features using the MS-FB and MS-SB approaches for the DBIL data. ($sb1$ is the lower sub-band and $sb2$ is the upper sub-band). 116
7.3 Baseline syllable recognition accuracy (SRA) and the conventional HMM (HTK) based syllable recognition accuracy (HTKSRA) of the joint features using the MS-FASB approach for the DBIL data. ($sb1$ is the lower sub-band and $sb2$ is the upper sub-band). 117

LIST OF FIGURES

1.1 Comparison of magnitude, wrapped phase, unwrapped phase and group delay spectra. (a) The z-plane with four complex poles and their complex conjugate pairs inside the unit circle, (b) the impulse response of the system shown in (a), (c) The magnitude spectrum of the signal shown in (b), (d) The wrapped phase spectrum of the signal shown in (b), (e) The unwrapped phase spectrum of the signal shown in (b), (f) the group delay function of the signal shown in (b). 4

2.1 Block diagram illustrating the three major operations in feature extraction for speech recognition. 12
2.2 The taxonomy of spectral analysis algorithms. 13
2.3 Block diagram illustrating the steps involved in the computation of the LPCC. 17
2.4 Block diagram illustrating the steps involved in the computation of the PLPCC. 18
2.5 Block diagram illustrating the steps involved in the computation of the Mel frequency cepstral co-efficients (MFCC) 21

3.1 Comparison of the minimum phase group delay function with the magnitude and linear prediction (LP) spectrum. (a) The z-plane with three poles inside the unit circle, (b) the magnitude spectrum of the system shown in (a), (c) the LPC spectrum of the system shown in (a), (d) the group delay spectrum of the system shown in (a). 29

3.2 Illustration of the manifestation of the additive property of the group delay function in resolving two closely spaced formants: z-plane, magnitude spectrum and group delay spectrum I) a pole inside the unit circle at $(0.8, \pi/8)$, II) a pole inside the unit circle at $(0.8, \pi/4)$ and III) a pole at $(0.8, \pi/8)$ and another pole at $(0.8, \pi/4)$, inside the unit circle. . . . 30

3.3 Comparison of the group delay spectrum with the FFT magnitude spectrum for a synthetic vowel. (a) The synthetic vowel with two resonant frequencies, (b) the FFT magnitude spectrum of the signal shown in (a), and (c) the group delay spectrum of the signal shown in (a). 32

4.1 Significance of proximity of zeros to the unit circle (a) The z-plane with four poles inside the unit circle, (b) the group delay spectrum of the system shown in (a), (c) The z-plane with four poles inside the unit circle and zeros added uniformly on the unit circle (d) the group delay spectrum of the system shown in (c), (e) The z-plane with zeros pushed radially inward into the unit circle, (f) the group delay spectrum of the system shown in (e). 35

4.2 Significance of pitch periodicity effects on the group delay function (a) The z-plane with three complex poles and their complex conjugate pairs inside the unit circle, (b) The impulse response of the system shown in (a), (c) The group delay spectrum of the signal shown in (b), (d) The response of the system shown in (a) to 5 impulses, and (e) The group delay spectrum of the signal shown in (d). 37

4.3 Comparison of various spectra for a synthetic signal (a) The synthetic signal with two resonances (b) the log magnitude spectrum of the signal shown in (a), (c) the root magnitude spectrum (root = 2/3) of the signal shown in (a), (d) the group delay spectrum of the signal shown in (a) and (e) the modified group delay spectrum of the signal shown in (a). . 41

5.1 Visualization of the correlation matrix of the MODGD feature set (measured on the NTIMIT database). 46

5.2 Comparison of the average error distributions (AED) of the MODGDF and root compressed cepstra in noise. (a) AED of the MODGDF (α = 0.4 , γ = 0.9) at 0 dB SNR, (b) AED of MODGDF at 3 dB SNR, (c) AED of MODGDF at 6 dB SNR, (d) AED of MODGDF at 10 dB SNR, (e) AED of the spectrally root compressed (SRC) cepstra (root = 2/3) at 0 dB SNR, (f) AED of SRC at 3 dB SNR, (g) AED of SRC at 6 dB SNR, (h) AED of SRC at 10 dB SNR, (i) AED of the energy root compressed (ERC) cepstra (root = 0.08) at 0 dB SNR, (j) AED of ERC at 3 dB SNR, (k) AED of ERC at 6 dB SNR, (l) AED of ERC at 10 dB SNR, (m) AED of the mel frequency (MFC) cepstra at 0 dB SNR, (n) AED of MFC at 3 dB SNR, (o) AED of MFC at 6 dB SNR, and (p) AED of MFC at 10 dB SNR. 48

5.3 Histograms of Bhattacharya distance criteria versus feature dimension for (a) Speaker separability, and (b) Language separability. 53

5.4 Two dimensional visualization of female-female speaker discrimination with the MODGDF using Sammon mapping. 54

5.5 Two dimensional visualization of English-French language discrimination with the MODGDF using Sammon mapping. 55

5.6 Results of Separability analysis. (a) Cumulative speaker separability of MODGDF and MFCC using Bhattacharya distance, and (b) Cumulative language separability of MODGDF and MFCC using Bhattacharya distance. 57

6.1 Comparison of the estimated flat spectrum for different cepstral window lengths. (a) & (b) A short segment of speech, (c) The squared magnitude spectrum $S(\omega)$ and its cepstrally smoothed version of $S_c(\omega)$ for a value of $lifter_w = 6$, (d) The squared magnitude spectrum $S(\omega)$ and its cepstrally smoothed version of $S_c(\omega)$ for a value of $lifter_w = 16$, (e) The estimated flat spectrum $E(\omega)$, for a value of $lifter_w = 6$, and (f) The estimated flat spectrum $E(\omega)$, for a value of $lifter_w = 16$. 65

6.2 Estimation of optimal $lifter_w$, α, and γ from a signal processing perspective. (a) z-plane plot of a system characterized by 3 formants (3 complex conjugate pole pairs), (b) Impulse response of the system shown in (a), (c) Response of the system in (a) excited with 5 impulses spaced 60 apart, (d) Group delay spectrum of the response in (a), (e) Group delay spectrum of the response in (c), (f) Modified group delay spectrum of the response in (c) for $lifter_w = 6$, $\alpha = 1$, and $\gamma = 1$, (g) Mean square error plot for α and γ (varied in steps of 0.1), (h) Modified group delay spectrum of the response in (d) for $lifter_w = 6$, $\alpha = 0.4$ and $\gamma = 0.9$. 67

6.3 English-French language discrimination with the MODGDF. 74

6.4 Korean-Japanese language discrimination with the MODGDF. 74

6.5 Comparison of segmentation of continuous speech using two methods. (a) Speech signal for Telugu speech utterance */mukhya mantri chandrabAbu nAyuDu AdEshham/*, (b) its short-time energy function, (c) minimum phase group delay function derived from short-time energy function. Thick vertical lines and dotted vertical lines denote the identified segment boundaries and the actual segment boundaries, respectively. Syllables at the bottom of (c) denotes the recognized syllables. . 76

6.6 Parallel network for a particular syllable string*/ni mukh yAm/* of Telugu language. 79

6.7 Reduced Vocabulary size Tamil and Telugu. (a) Tamil and (b) Telugu. 83

6.8 Spectrogram like plots to illustrate formant reconstructions for a synthetic vowel. (a) The short-time modified group delay spectra reconstructed from MODGDF (b) The short-time power spectra reconstructed from MFCC, (c) The short-time composite spectra reconstructed from joint features (MODGDF+MFCC), and (d) The short-time power spectra reconstructed from RASTA filtered MFCC. 90

6.9 Spectrogram like plots to illustrate formant reconstructions for a synthetic speech signal with varying formant trajectory. a) The short-time modified group delay spectra reconstructed from the MODGDF b) The short-time power spectra reconstructed from MFCC, and c) The short-time composite spectra reconstructed from joint features (MODGDF+MFCC) 91

6.10 Spectrogram like plots to illustrate formant reconstructions for clean and noisy speech. (a) The reconstructed short-time modified group delay spectra for clean speech, (b) The reconstructed short-time modified group delay spectra for the same segment of speech corrupted with car noise at a SNR of 2 dB, (c) The reconstructed short time power spectra (MFCC) for clean speech, (d) The reconstructed short time power spectra (MFCC) for the same segment of speech corrupted with car noise at a SNR of 2 dB, (e) The reconstructed short time composite spectra from joint features (MODGDF+MFCC) for clean speech, (f) The reconstructed short time composite spectra from joint features (MODGDF+MFCC) for the same segment of speech corrupted with car noise at a SNR of 2 dB. 92

7.1 Block diagram of the MS paradigm used in speech recognition. 105

7.2 Block diagram of the MB paradigm used in speech recognition. 106

7.3 Block diagram of the single-stream (SS-FB) paradigm 108

7.4 Block diagram of the MS-FB paradigm 109

7.5 Response of the two triangular filters used to derive the sub-bands in the SS-MB approach 110
7.6 Block diagram of the SS-MB paradigm 110
7.7 Block diagram of the MS-SB paradigm 112
7.8 Block diagram of the MS-FASB paradigm 113

B.1 Partitioning of the 2 dimensional sample space into 15 cells using vector quantization. 126
D.2 Topology of a 3 state left to right continuous HMM with three mixtures per state. 133

ABBREVIATIONS

AED - Average Error Distribution
ASR - Automatic Speech Recognition
CMS - Cepstral Mean Subtraction
DARPA - Defense Advanced Research Project Agency
DBIL - Database of Indian Languages
DCT - Discrete Cosine Transform
DFT - Discrete Fourier Transform
EM - Expectation Maximization
ERMFC - Energy Root Compressed Mel-Frequency Cepstral Coefficients
FFT - Fast Fourier Transform
GD - Group Delay
GMM - Gaussian-Mixture Model
HMM - Hidden Markov Model
LBG - Linde Buzo Gray
LFCC - Linear Frequency Cepstral Coefficients
$lifter_w$ - Cepstral Window
LPCC - Linear Prediction Cepstral Coefficients
LPC - Linear Prediction Coefficients
LP - Linear Prediction
MAP - Maximum A posteriori Probability
MB - Multi-Band
MB-SS - Multi-Band Single-Stream
MFCC - Mel-Frequency Cepstral Coefficients
ML - Maximum Likelihood
MODGDF - Modified Group Delay Feature
MODGD - Modified Group Delay

MSE - Mean Squared Error
MS-FA-SB - Multi-Stream Full and Selective-band
MS-FB - Multi-Stream Full-Band
MS - Multi-Stream
MS-SB - Multi-Stream and Selective-band
OGI - Oregon Graduate Institute
OGI_MLTS - Oregon Graduate Institute Multi-language Telephone Speech
PDF - Probability Density Function
PLP - Perceptual Linear Prediction
RASTA - RelAtive SpecTrAl
RMS - Root Mean Square
SFS - Sequential Forward Search
SNR - Signal to Noise Ratio
SPINE - Speech in Noisy Environments
SRI - Stanford Research Institute
SRLFC - Spectral Root Compressed Linear Frequency Cepstral Coefficients
SRMFC - Spectral Root Compressed Mel-Frequency Cepstral Coefficients
SS-FB - Single-Stream Full-Band
TIMIT - Texas Instruments and Massachusetts Institute of Technology
VQ - Vector Quantization

CHAPTER 1

INTRODUCTION

Human perception involves both classification and recognition. The advent of language is perhaps a good example of the human disposition to inherently classify and recognize patterns. Discovering and recognizing patterns present in the speech signal is probably the most difficult task in pattern recognition by machines. Feature extraction is the task of converting a sequence of speech samples at the front end into a set of observation vectors which represent events in a probabilistic space over which classification is performed. The area of feature extraction for speech recognition is a challenging area. Features that are robust and capture both absolute and dynamic spectral information in the speech signal are often difficult to extract. The merger of the feature extraction problem with the speech recognition process, to estimate more complex models in a closed loop manner, makes the area more challenging. Hence there is a need to explore new methods of feature extraction for speech recognition.

1.1 MOTIVATION

Spectral representation of speech is complete when both the Fourier transform magnitude and phase spectra are specified. In conventional speech recognition systems, features are generally derived from the short time magnitude spectrum. Although the importance of Fourier transform phase in speech perception has been realized [1], few attempts have been made to extract features from it. This is primarily because the resonances of the speech signal which manifest as transitions in the phase spectrum are completely masked by the wrapping of the phase spectrum. Hence an alternative to processing the Fourier transform phase, for extracting speech features, is to process the group delay function which can be directly computed from the speech signal

[2]. The group delay function has been used in earlier efforts [8], to extract pitch and formant information from the speech signal. In all these efforts, no attempt was made to extract features from the speech signal and use them for speech recognition applications. Hence the focus of this thesis is to formulate and investigate methods to extract speech features from the modified group delay function, which is based on the Fourier transform phase and to use them for various speech recognition applications.

1.2 IMPORTANCE OF PHASE IN SPEECH PROCESSING

Short time Fourier analysis can be used to process speech assuming that it is quasi stationary after applying a window on the speech signal. Let $x(n)$ be a given speech sequence and $X_n(\omega)$ its short time Fourier transform (STFT), after applying a window $w(n)$, on the speech signal $x(n)$

$$X_n(\omega) = \sum_{m=-\infty}^{+\infty} x(m)w(n-m)e^{j\omega m} \tag{1.1}$$

The STFT $X_n(\omega)$ can also be expressed as

$$X_n(\omega) = |X_n(\omega)|e^{j\theta_n(\omega)} \tag{1.2}$$

In Equation 1.1, n represents the time instant at which the Fourier transform is evaluated. In Equation 1.2, $|X_n(\omega)|$ corresponds to the short time magnitude spectrum and $\theta_n(\omega)$ corresponds to the phase spectrum. The square of the short-time magnitude spectrum $|X_n(\omega)|^2$ is called the short time power spectrum. The speech signal is therefore completely characterized by both the magnitude and phase spectrum. Speech is generally represented by using the transform domain approach, followed by spectral compression and homomorphic processing. The most successful approach to spectral representation of speech involves passing the short-time power spectra through a set of non linear filter banks followed by the extraction of cepstral features. The contemporary features used in speech recognition can be classified as

- Articulatory or model based features like Linear prediction cepstral coefficients and line spectral pairs.

- Perceptual or spectral features like Mel frequency cepstral coefficients and perceptual critical bank features.
- Prosodic or Long-term features like energy, power and F0 contours.
- Source features derived from the linear prediction residual.

Spectral features are generally derived from the short-time Fourier transform magnitude spectrum. The short-time phase spectrum is not utilized in most of the conventional feature extraction techniques. The two primary reasons for this are

- The Fourier transform phase spectrum is generally available in a wrapped form. The discontinuities in the phase spectrum often called phase wrapping, have to be corrected by a non-unique process of phase unwrapping.
- Early psychoacoustic experiments conducted by Drucker [9] and Helmholtz [10] have claimed that the human perception mechanism is not sensitive to phase.

But the importance of Fourier transform phase in speech perception has been realized in [1], where it was shown that human perception mechanism is sensitive to phase of intervocalic plosives. Recent experiments conducted by Paliwal [11], have also brought out the significance of the short-time phase spectrum in speech perception and recognition. The phase spectrum has also been used in speech processing by Yegnanarayana and Murthy [12] via the group delay domain. Although the group delay function is defined as the negative derivative of the unwrapped phase spectrum, it can be computed directly from the speech signal as in [2]. Hence the phase spectrum can be utilized via the group delay domain without having to unwrap the phase spectrum. To illustrate this we consider a four formant system characterized by four complex poles with their complex conjugate pairs shown in Figure 1.1 (a). The impulse response of such a system is shown in Figure 1.1 (b). The magnitude spectrum of the signal shown in Figure 1.1 (b) is shown in Figure 1.1 (c). The resonances of the signal manifest as peaks in the magnitude spectrum. The phase spectrum in a wrapped form is shown in Figure 1.1 (d). The corresponding unwrapped phase spectrum is shown in Figure 1.1 (e). While the resonances of the signal are not visible in the wrapped phase spectrum, they manifest as transitions in the unwrapped phase spectrum. The transitions in the unwrapped phase spectrum corresponding to the resonances are not clear. The group delay function computed from the signal in Figure 1.1 (b) is shown in Figure 1.1 (f). The resonances of the signal are clearly perceived in the group delay domain. Hence an alternative to processing the short-time phase spectrum is to process the group delay function computed directly from the speech signal. The group delay function has been

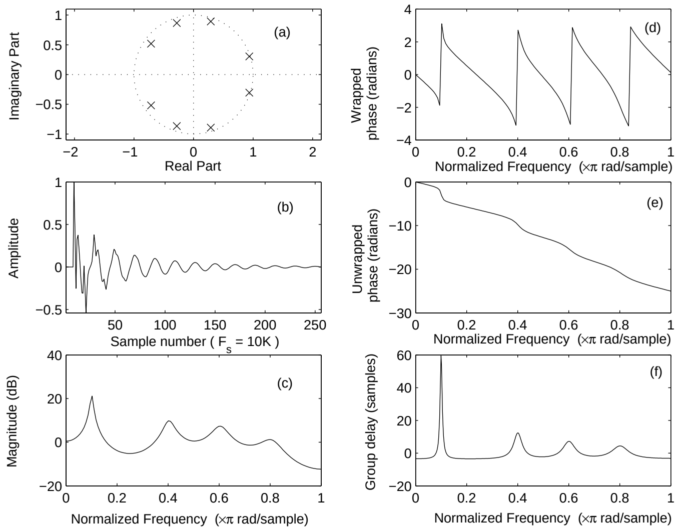

Fig. 1.1: Comparison of magnitude, wrapped phase, unwrapped phase and group delay spectra. (a) The *z*-plane with four complex poles and their complex conjugate pairs inside the unit circle, (b) the impulse response of the system shown in (a), (c) The magnitude spectrum of the signal shown in (b), (d) The wrapped phase spectrum of the signal shown in (b), (e) The unwrapped phase spectrum of the signal shown in (b), (f) the group delay function of the signal shown in (b).

used in earlier efforts for pitch and formant extraction in [8] and spectrum estimation in [13]. No attempt was made in all these efforts to use the group delay function to extract speech features and use them for speech recognition applications.

1.3 OBJECTIVES OF THE THESIS

The following are the main objectives of the thesis:

- To propose a new representation of speech derived from the modified group delay function which is based on the Fourier transform phase.
- To evaluate the feature using standard feature evaluation criteria.
- To analyze the complementary nature of the feature derived from the modified group delay function with respect to features derived from the short-time magnitude spectrum.
- To investigate the significance of the new feature in the multi-stream and multi-band paradigm of automatic speech recognition.

1.4 ORGANIZATION OF THE THESIS

Chapter 2 gives a background to feature extraction for speech recognition. The various approaches used for feature extraction for speech recognition are discussed here. This is followed by a discussion of the contemporary features used in speech processing. This Chapter concludes with a brief review of the three most commonly used features in speech recognition namely, the perceptual linear prediction co-efficients (PLPCC), Relative-Spectral perceptual linear prediction co-efficients (RASTA-PLP), and the Mel frequency cepstral co-efficients (MFCC).

The theory of group delay functions is introduced in Chapter 3. The Fourier transform phase spectrum is available in a wrapped form. The phase spectrum has to be unwrapped for any meaningful use in speech recognition applications. The Group delay function is defined as the negative derivative of the unwrapped phase spectrum and has been extensively used for speech processing. The advantage of the group delay

function instead of the phase spectrum is that it can computed from the signal itself. Hence the problem of unwrapping the phase spectrum can be avoided. The theory and issues in the computation of the group delay function are discussed first. The relation between the minimum phase group delay function (group delay spectrum) and the magnitude spectrum is described next. The two properties of the group delay function namely the additive property, and the high resolution property which are of relevance to this thesis are also discussed here.

The group delay function becomes spiky and distorted due to zeros that are very close to the unit circle in the z-plane and also due to pitch periodicity effects. The dynamic range of the group delay spectrum is also considerably altered due to these effects. The group delay function has to be modified to suppress these effects. Chapter 4 discusses the basis for modifying the group delay function. The zeros that lie very close to the unit circle in the z-plane are pushed radially into the unit circle by replacing the denominator of the group delay function with its cepstrally smoothed version. It is also illustrated how such a modification results in restoring the formant structure of the speech signal. Two new parameters α and γ are also introduced to reduce the dynamic range of the modified group delay spectrum. Cepstral features are extracted from the modified group delay function. The second form of the discrete cosine transform DCT-II is used transform the modified group delay function to the modified group delay cepstral co-efficients. The modified group delay cepstral co-efficients are called the modified group delay feature (MODGDF).

The MODGDF is evaluated using several feature evaluation criteria in Chapter 5. A measured correlation matrix is visualized to show that the MODGDF is indeed decorrelated. The robustness of the MODGDF to convolutional and white noise is analyzed next. It's similarity to RASTA and the significance of cepstral mean subtraction in the group delay domain are also discussed in this Chapter. The task independence of the MODGDF is illustrated using the sequential forward search (SFS) feature selection method. The results of two dimensional cluster analysis are presented. The results indicate that the MODGDF is able to discriminate speakers and languages linearly, when they are considered on a pairwise basis. The results of cumulative separability analysis in the higher dimensional feature space using the Bhattacharya distance also indicate that the MODGDF performs relatively better than the MFCC.

Chapter 6 evaluates the performance of the MODGDF for three speech recognition tasks namely speaker identification, language identification, and continuous speech recognition. Results of automatic speaker identification for clean speech (on the TIMIT database) and noisy telephone speech (on the NTIMIT database) using the MODGDF are presented. This is followed by a description of experimental results of automatic language identification using the MODGDF for the 3 language task (on the DBIL and OGI_MLTS database) and the 11 language task (on the OGI_MLTS database). The performance of the MODGDF for syllable based continuous speech recognition is discussed next. Broadcast news corpora from the DBIL database [5] of two Indian languages Tamil and Telugu are used for this task. The baseline recognition system uses Hidden Markov Models trained apriori for 320 syllables for Tamil and 265 syllables for Telugu. The number of syllables used for training are selected based on their frequency of occurrence in the respective corpora. Any syllable that occurs more than 50 times in the corpus is selected as a candidate for which HMMs are built. A separate model is built for silence. The test phrase is segmented at boundaries of syllabic units using the minimum phase group delay function derived from the short-time energy function [14, 15]. These segments are now checked in isolated style against all HMMs built apriori using the reduced vocabulary. The HMM that gives the maximum likelihood value is declared as the correct match. Finally the output syllable sequence is generated by concatenating all the isolated syllables, preserving the order of the original phrase. An enhanced baseline system which uses local forced Viterbi realignment is also described. The experimental results using the MODGDF for the continuous speech recognition task are discussed. The results of performance evaluation of the MODGDF for all the three speech recognition tasks are compared to the performance of the log compressed and root compressed forms of the MFCC and LFCC.

The significance of the MODGDF in the multi-stream and multi-band paradigm of automatic speech recognition is discussed in Chapter 7. The Chapter begins with a brief description of some benchmark multi-stream and multi-band approaches. Conventional multi-stream and multi-band approaches combine multiple streams of features derived from narrow frequency bands over the entire frequency band using early fusion or late fusion. Recombination at the feature level which is often called early fusion is performed by a simple concatenation of multiple feature streams across multiple frequency bands over the entire frequency band. In this Chapter, a new approach within the multi-stream and multi-band framework for ASR is proposed. The whole frequency band is divided into multiple frequency bands. The MODGDF which is based on the Fourier transform phase and the MFCC derived from the Fourier transform magnitude are extracted across these multiple frequency bands. The two feature streams are then recombined across selective frequency bands using early fusion. This approach is called the multi-stream and selective-band recombination (MS-SB) approach. This approach is also extended to combining features extracted from the entire frequency band with the features extracted from the MS-SB approach. This approach is called the multi stream with full and selective band (MS-FA-SB) approach. A syllable based continuous speech recognition system for two Indian languages Tamil and Telugu is implemented using the MS-SB and the MS-FA-SB approaches. Significant improvements in recognition performance in the range of 13-14% are achieved over the conventional single-stream full-band (SS-FB) approach.

A summary of the Thesis is given in Chapter 8. The significance of the MODGDF as a new feature for speech recognition is discussed. The criticisms on the work presented in the thesis are also listed. The Chapter concludes with a discussion on the scope for future work.

1.5 MAJOR CONTRIBUTIONS OF THE THESIS

- A new representation of speech called the modified group delay feature, which is based on the Fourier transform phase is proposed for speech recognition applications.

- The MODGDF is used in practice across three speech recognition tasks namely automatic speaker identification, automatic language identification, and syllable based continuous speech recognition.
- The complementary nature of the MODGDF derived from the modified group delay function with respect to features derived from the Fourier transform power spectra (MFCC) is illustrated with the help of extensive experiments.
- The significance of the MODGDF in the multi-stream and multi-band paradigm of automatic speech recognition is highlighted.
- A new approach called the multi-stream and selective band (MS-SB) approach is proposed within the framework of the multi-stream and multi-band paradigm, for improved speech recognition performance.

CHAPTER 2

BACKGROUND TO SPEECH FEATURE EXTRACTION

The role of the human speech production and perception mechanism is very significant in the study of the different methods of speech feature extraction. Speech is produced due to the vibration of the vocal cords (vocalics) when breath is exhaled from the lungs. The variety in speech sounds is primarily due to the variation in mass and length of the vocal cords and the articulators of the vocal tract. Men have vocal cords of length 17 - 24 mm in length and an average fundamental frequency of 125 Hz, while women have vocal cords of length 13 - 17 mm and an average fundamental frequency of 200 Hz. Although it is not clear how the brain interprets speech signals, it is well known that the human ears sensitivity to speech is non linear [16]. A non uniform scale based on the Mel which is defined as one thousandth of the pitch of a 1000 Hz tone [17], is usually used to simulate this non uniform sensitivity of the ear for the purpose of speech feature extraction.

2.1 FEATURE EXTRACTION VERSUS CLASSIFICATION

Feature Extraction is a process of converting a sequence of speech samples into a set of observation vectors which represent events in a probabilistic space over which classification is performed. Feature extraction is also defined as mapping of an n dimensional vector, x, to a vector y, of dimension m where $m < n$ such that an appropriate criterion is optimized. Classification is defined as mapping an observation vector x to a class ω_i, where $i = 1, 2, 3, ..., L$ and L is the number of classes. There exists a high degree of correlation between feature extraction methods and the design of classifiers for speech recognition. Depending on the classification problem at hand, the design complexity of a speech recognition system is distributed between the feature extraction step and the classification step. An efficient classifier distributes this complexity between the feature extraction step and the classification step appropriately. Hence the feature extraction step shares the overall complexity of implementation of a speech recognition system. In this context, the three major challenges in designing feature extraction systems are

1. To extract perceptually meaningful features.
2. To extract robust or invariant features.
3. To extract features that capture the temporal correlation of the spectrum.

2.2 THE SPEECH FEATURE EXTRACTION PARADIGM

In this Section, the three major steps in the extraction of features from the speech signal are reviewed. A typical speech feature extraction system involves three basic steps

1. Spectral shaping.
2. Spectral analysis.
3. Parametric transformation.

The block diagram illustrating the three basic steps in feature extraction for speech recognition is shown in Figure 2.1. These three steps are reviewed briefly in the following Sections.

2.2.1 Spectral Shaping

The natural speech signal is available in an analog form. Spectral shaping of the natural speech signal begins with an analog to digital conversion of the signal. The microphone used to acquire the speech signal introduces nonlinear distortion and line frequency noise. Further the A-D conversion process introduces its own distortion due to the non linear transfer function of the A-D converter. Sampling rates of 8, 10, 12, and 16 K are generally used to digitize the speech signal [18]. These sampling rates

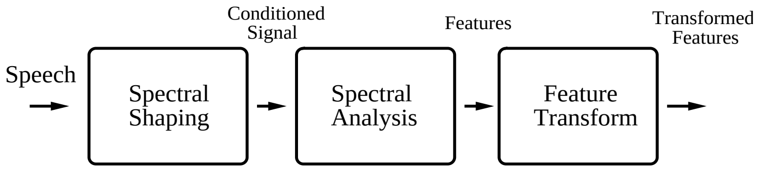

Fig. 2.1: Block diagram illustrating the three major operations in feature extraction for speech recognition.

give good time and frequency resolution. The digitized signal is passed through a finite impulse response (FIR) filter with a transfer function

$$H(z) = \sum_{k=0}^{k=N-1} a(k) z^{-k} \tag{2.1}$$

where N is the length of the filter.
In practice a one co-efficient digital filter given by

$$H(z) = 1 + az^{-1} \tag{2.2}$$

called the pre-emphasis filter is used. The value of a generally varies from -1.0 to -0.4. Natural speech has a spectral tilt of - 20 dB/octave. The pre-emphasis filter is used to offset this spectral tilt of natural speech. Spectral shaping is generally followed by spectral analysis.

2.2.2 Spectral Analysis

Spectral analysis algorithms can be classified into six major classes [18]. This taxonomy is illustrated in Figure 2.2. Early spectral analysis methods computed filter bank energies after passing the speech spectrum through a filter bank. Linear prediction based techniques were extensively used in the decade from 1970 to 1980. Fourier transform techniques have been used widely thereon, although variations of linear prediction techniques like the perceptual linear prediction (PLP) continue to be used. The natural acoustic frequency f is mapped to a non-linear perceptual frequency scale

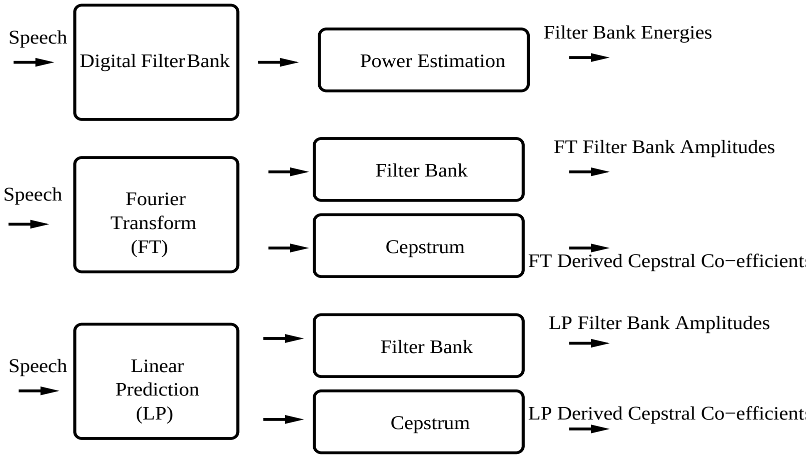

Fig. 2.2: The taxonomy of spectral analysis algorithms.

which is generally the *Bark* or the *Mel* scale. The *Bark* scale is defined as

$$B_f = 13\ tan^{-1}\left(\frac{0.76f}{1000}\right) + 3.5\ tan^{-1}\left(\frac{f^2}{(7500)^2}\right) \tag{2.3}$$

where B_f is the *Bark* frequency scale.
The *Mel* scale is defined as

$$M_f = 2595\ log_{10}\left(1 + f/700\right) \tag{2.4}$$

where M_f is the *Mel* frequency scale. The natural acoustic frequency f is mapped to a non-linear perceptual frequency scale by passing the speech spectrum through a filter bank placed on the *Bark* or the *Mel* scale. Cepstral features used in speech recognition are based on homomorphic signal processing techniques. They are significant in the context of speech feature extraction due to their ability to separate the source from the system (deconvolution). Cepstral co-efficients derived from the Fourier transform filter bank energies are called Mel frequency cepstral co-efficients (MFCC). Linear prediction (LP) co-efficients are extracted using parametric modeling techniques which model

the speech spectrum as an autoregressive process. They can either be derived as LP filter bank amplitudes which result from sampling the LP spectrum at critical filter bank frequencies or the LP cepstral co-efficients which are the cepstral co-efficients computed from the LP spectrum. Cepstral co-efficients derived from the critical band LP spectrum are called the perceptual linear prediction co-efficients (PLPCC). Since spectral analysis results in a set of correlated features which can capture only the static information in the speech signal, it is generally followed by a feature transformation step.

2.2.3 Feature Transformation

The speech signal is characterized by temporal variations which are difficult to characterize with absolute measurements. Hence higher order time derivatives are often appended to the absolute feature vectors. The first and the second order time derivatives of the absolute feature vectors are commonly used in speech recognition. The first and the second order time derivatives are computed by using either the forward or the backward difference between adjacent frames of speech using regression analysis. In the light of the discussion so far, a brief review of the approaches to speech feature extraction is given in the following Section.

2.3 REVIEW OF APPROACHES TO SPEECH FEATURE EXTRACTION

Speech feature extraction techniques can be classified under the following broad categories

1. *Waveform domain approach:* This approach uses amplitude, duration, zero crossing rate and level crossing rate of the speech signal as features. Sheikhzadeh and Deng [19], used a waveform based recognition system that combines a time varying autoregressive filter with a hidden Markov model. The disadvantages of this approach are the high data rate involved and the evidence that the ear itself performs some kind of signal transformation.

2. *Auditory spectral model based approach:* This approach involves the explicit simulation of the human auditory model in terms of firing activity in the cochlea. The most popular models are the Ensemble Interval Histogram (EIH) model by Ghitza [20], the Payton model [21], the cochlear auditory model by Lyon [22] and the Seneff auditory model [23]. The disadvantages are the relatively poor recognition performance and lack of understanding the role of the brain in human perception.
3. *Transform domain approach:* Speech features are extracted in the transformed domain in this approach, capturing relevant information and discarding redundant information from the speech signal. Some of the popular transform domain methods are the filter banks, discrete Fourier transform (DFT), linear prediction co-efficients (LPC) and orthogonal transforms like the Karhunen-Loeve transform (KLT) and the discrete cosine transform (DCT).
4. *Homomorphic signal processing approach:* This is the most popular and widely used signal processing approach for speech feature extraction. Bogert, Healy and Tukey [24] used the homomorphic signal processing techniques to process seismic signals. Noll [25] applied the cepstral notions to speech processing for the first time. Various researchers like Rabiner, Juang [26], and Quatieri [27] have pioneered the work on the use of cepstral features for speech recognition.
5. *Spectral compression approach:* Early investigations by Stevens and Volkman on the human hearing system [16] introduced the Mel spectrum. Later it was successfully used by Fant [28], Picone [18], Mermelstein and Davis [17] to extract features from the speech signal for improved recognition performance. Stephens and Bate [29] justified the Mel spectrum in terms of the auditory model.
6. *Joint time frequency domain approach:* Spectrograms were the first kind of time frequency distributions (TFD) used for signal modeling. Pitton, Atlas and Loughlin [30] investigated TFDs for speech processing. Reily and Boashash have compared TFDs with wavelets [31], while Rainton and Young [32], have

compared TFDs with filter bank energies as speech feature vectors.

7. *Spectral shaping approach:* Voiced speech has a negative spectral slope or tilt of -20 dB/octave due to physiological characteristics of the speech production mechanism. The spectral slope can be included in the speech representation process either by using a pre-emphasis filter or by adding power to some coefficients directly on the spectrum. MIT Lincoln labs [33], have used features computed using this approach along with other features for improving speech recognition performance. Murthy, Beaufays, and Heck [34], have used a perceptually motivated feature like the Mel slope for speaker identification over telephone channels.

In the following Section, some of the most commonly used speech feature extraction techniques are detailed.

2.4 SOME POPULAR FEATURE EXTRACTION TECHNIQUES

Several techniques have been used to extract features from speech [18]. With the advent of Markov models in speech recognition, features that are perceptually meaningful and invariant to the ambient acoustic environment have become increasingly common. The contemporary features used in speech recognition are the Mel frequency cepstral coefficients (MFCC), perceptual linear prediction cepstral coefficients (PLPCC), perceptual critical bank features (PCBF), energy, power, F0 contours and source features. The MFCC are perhaps the most commonly used features in automatic speech recognition. Spectral features like the MFCC are generally computed from the short-time Fourier transform magnitude spectrum. Model based features like the LPCC and the PLPCC are computed using linear prediction analysis. In this Section, computation of the LPCC, PLPCC, LFCC and MFCC are discussed briefly.

2.4.1 Linear Prediction Cepstral Co-efficients

The linear prediction cepstral co-efficients (LPCC) used extensively for speech recognitionin in the decade from 1970 to 1980, belong to the generic model based approaches. Makhoul [35] introduced linear prediction methods for speech recognition. The block diagram for computing LPCC is given in Figure 2.3. The steps involved in the computation of the LPCC are

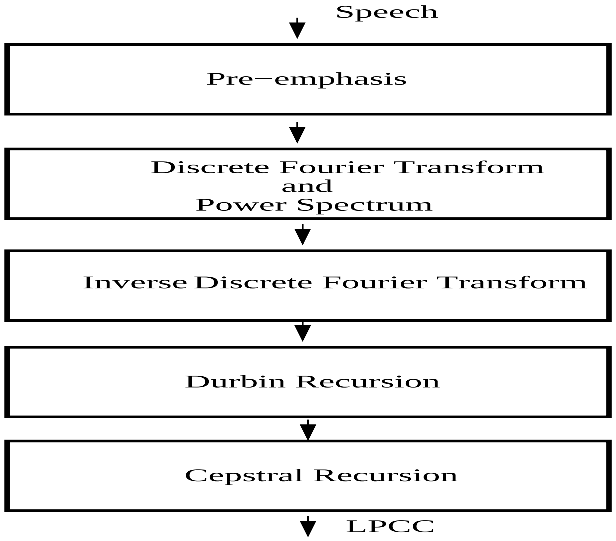

Fig. 2.3: Block diagram illustrating the steps involved in the computation of the LPCC.

- Perform frame blocking and windowing on the speech signal.
- Compute the discrete Fourier transform (DFT) and its squared magnitude.
- Perform an inverse discrete Fourier transform (IDFT).
- Derive an autoregressive model using regression analysis.

- Use an orthogonal transformation like the KLT or the DCT to compute decorrelated LPCC.

2.4.2 Perceptual Linear Prediction Co-efficients

The perceptual linear prediction (PLP) maps the linear prediction (LP) spectrum to the nonlinear frequency scale of the human ear. The perceptual linear prediction co-efficients (PLPCC) are an extension of the LPCC. The PLPCC give a relatively better speech recognition performance when compared to the LPCC. A block diagram illustrating the steps involved in the computation of the PLPCC is given in Figure 2.4. The steps involved in the computation of the PLPCC are

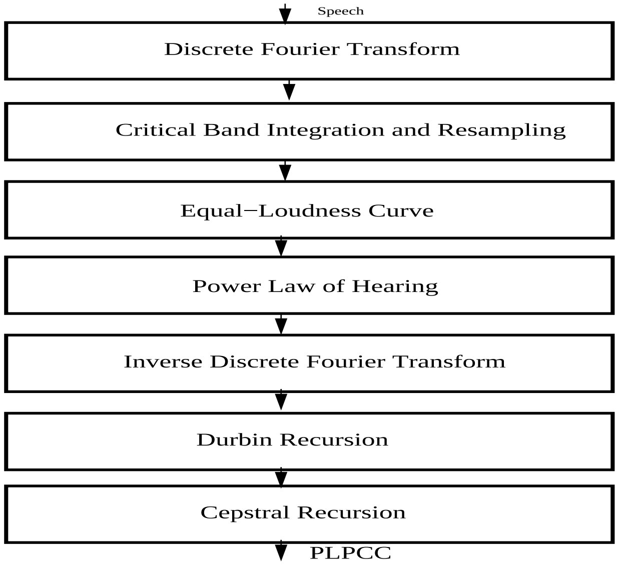
Speech
Discrete Fourier Transform
Critical Band Integration and Resampling
Equal−Loudness Curve
Power Law of Hearing
Inverse Discrete Fourier Transform
Durbin Recursion
Cepstral Recursion
PLPCC

Fig. 2.4: Block diagram illustrating the steps involved in the computation of the PLPCC.

- Perform frame blocking and windowing on the speech signal.

- Compute the discrete Fourier transform (DFT) and its squared magnitude.
- Integrate the power spectrum hence computed within overlapping critical band filter responses.
- Pre-emphasize the spectrum to simulate the unequal sensitivity of the human ear to different frequencies.
- Compress the spectral amplitudes by taking the cube root after integration.
- Perform an inverse discrete Fourier transform (IDFT).
- Perform spectral smoothing on the critical band spectra using an autoregressive model derived from regression analysis.
- Use an orthogonal transformation like the KLT or the DCT to compute decorrelated PLPCC.
- Optionally liftering can be performed to equalize the variances of the different cepstral co-efficients.

2.4.3 Linear Frequency Cepstral Co-efficients

The linear frequency cepstral co-efficients (LFCC) were cepstral features that were used prior to the advent of the MFCC. The LFCC are computed as

- The speech signal x(n) is windowed with an analysis window w(n)and its discrete short time Fourier transform $X(n, \omega_k)$ is computed as

  $$X(n,\omega_k) = \sum_{m=-\infty}^{+\infty} x(m)w(n-m)e^{-j\omega_k m} \tag{2.5}$$

  where

  $$\omega_k = \frac{2\pi}{N}k \tag{2.6}$$

  where N is the DFT length, and n, k are integers.

- The square of the magnitude of $X(n, \omega_k)$, is then weighted by a series of filter frequency responses distributed over the linear scale.

- The linear frequency cepstral co-efficients are computed as

  $$C_{lin}[n,m] = \frac{1}{R}\sum_{l=0}^{R-1} log\{E_{lin}(n,l)\}cos(\frac{2\pi}{R}lm) \tag{2.7}$$

  where R is the number of filters and the inverse transform is in terms of the discrete cosine transform (DCT) and $E_{lin}(n,l)$ is the energy in each frame of the speech signal at time n and for the l^{th} linear scale filter.

The LFCC are also computed by directly taking the discrete cosine transform (DCT) of the log-magnitude discrete Fourier transform spectrum [17].

2.4.4 Mel Frequency Cepstral Co-efficients

The Mel frequency cepstral co-efficients (MFCC) are perhaps the most widely used features in speech recognition today. Stevens and Volkman [16], developed the Mel scale as a result of a study of the human auditory perception. The Mel scale was used by Mermelstein and Davis [17], to extract features from the speech signal for improved recognition performance. The block diagram for the computation of the MFCC is shown in Figure 2.5. The steps followed in the computation of the MFCC are

- The speech signal x(n) is windowed with an analysis window w(n)and its discrete short time Fourier transform $X(n, \omega_k)$ is computed as

  $$X(n,\omega_k) = \sum_{m=-\infty}^{+\infty} x(m)w(n-m)e^{-j\omega_k m} \tag{2.8}$$

  where

  $$\omega_k = \frac{2\pi}{N}k \tag{2.9}$$

  where N is the DFT length, and n, k are integers.

- The square of the magnitude of $X(n, \omega_k)$, is then weighted by a series of filter frequency responses distributed over the Mel scale.
- The energy in each frame of the speech signal at time n and for the l^{th} Mel scale filter is now computed using

  $$E_{Mel}(n,l) = \frac{1}{A_l}\sum_{k=L_l}^{U_l} |V_l(\omega_k)X(n,\omega_k)|^2 \tag{2.10}$$

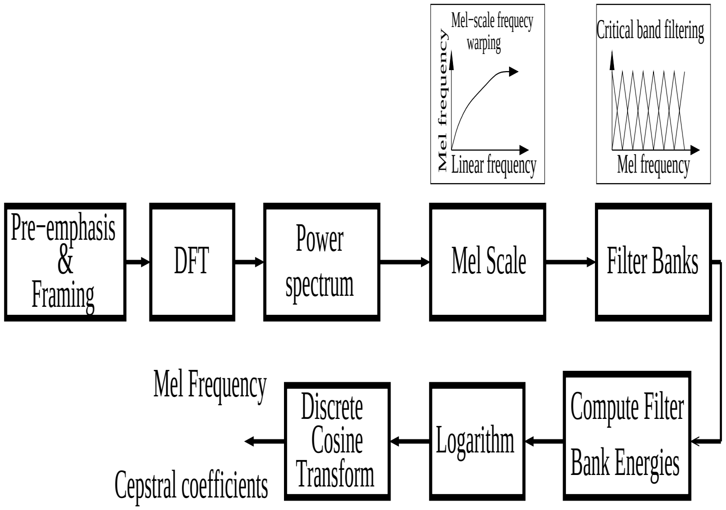

Fig. 2.5: Block diagram illustrating the steps involved in the computation of the Mel frequency cepstral co-efficients (MFCC)

where L_l and U_l are the upper and the lower cutoffs of each filter.

A_l is given by:

$$A_l = \sum_{k=L_l}^{U_l} |V_l(\omega_k)|^2 \tag{2.11}$$

- Finally the Mel frequency cepstral co-efficients are computed as

$$C_{Mel}[n,m] = \frac{1}{R}\sum_{l=0}^{R-1} log\{E_{Mel}(n,l)\}cos(\frac{2\pi}{R}lm) \tag{2.12}$$

where R is the number of filters and the inverse transform is in terms of the discrete cosine transform (DCT).

- The use of the DCT results in a set of decorrelated vectors which enables the use of diagonal covariances in modeling the feature vector distribution.

2.5 SUMMARY

This Chapter gives a brief background to speech feature extraction, relevant to the work presented in the thesis. Feature extraction is a process of extracting meaningful features from the speech signal. Various techniques have evolved over the years. The significance of feature extraction and its relation to classification techniques is discussed first. This is followed by a brief review of the basic steps used in speech feature extraction. The various approaches used in feature extraction for speech recognition is discussed next. Computation of the most commonly used features like the linear prediction co-efficients, perceptual linear prediction co-efficients, and the Mel frequency cepstral co-efficients are discussed.

CHAPTER 3

THEORY AND PROPERTIES OF GROUP DELAY FUNCTIONS

The characteristics of the speech signal can be visually perceived in the short time magnitude spectra in comparison to the short time phase spectra. With specific reference to the speech signal it can be stated that the resonances of the speech signal present themselves as the peaks of the envelope of the short time magnitude spectrum. These resonances often called formants manifest as transitions in the short time phase spectrum. The problem with identifying these transitions is the masking of these transitions due to the wrapping of the short-time phase spectrum at multiples of 2π. Hence any meaningful use of the short time phase spectrum for speech processing involves the non-unique process of phase unwrapping. The group delay function which is defined as the negative derivative of the unwrapped short-time phase spectrum, can be computed directly from the speech signal as in [2], without unwrapping the short-time phase spectrum. The group delay function has been effectively used to extract various source and system parameters [12], when the signal under consideration is a minimum phase signal. This is primarily because the magnitude spectrum of a minimum phase signal [12], and its group delay function resemble each other [36].

3.1 THE GROUP DELAY FUNCTION

Group delay is defined as the negative derivative of the unwrapped Fourier transform phase. Mathematically the group delay function is defined as

$$\tau(\omega) = -\frac{d(\theta(\omega))}{d\omega} \tag{3.1}$$

where the phase spectrum $(\theta(\omega))$ of a signal is defined as a continuous function of ω. The values of group delay function away from a constant indicates the degree of

nonlinearity of the phase. The group delay function can also be computed from the signal as in [37], using

$$\tau_x(\omega) = -Im\frac{d(log(X(\omega)))}{d\omega} \tag{3.2}$$

$$= \frac{X_R(\omega)Y_R(\omega) + Y_I(\omega)X_I(\omega)}{|X(\omega)|^2} \tag{3.3}$$

where the subscripts R and I, denote the real and imaginary parts of the Fourier transform. $X(\omega)$ and $Y(\omega)$ are the Fourier transforms of $x(n)$ and $nx(n)$, respectively. It is also important to note in this context that the group delay function can be expressed in terms of the cepstral co-efficients as

$$\tau(\omega) = -\theta^{'}(\omega) = \sum_{n=1}^{\infty} nc(n)\ cos(n\omega) \tag{3.4}$$

where $c(n)$ are the n dimensional cepstral co-efficients. Hence the group delay function $\tau(\omega)$, in general, can also be viewed as the Fourier transform of the weighted cepstrum.

3.2 GROUP DELAY SPECTRUM AND MAGNITUDE SPECTRUM

In general, if we consider the spectrum of any signal as a cascaded of M resonators, the frequency response of the overall filter is given by [36]

$$X(e^{j\omega}) = \prod_{i=1}^{M} \frac{1}{\alpha_i^2 + \beta_i^2 - \omega^2 - 2j\omega\alpha_i} \tag{3.5}$$

where $\alpha_i \pm j\beta_i$ is the complex pair of poles of the i^{th} resonator. The squared magnitude spectrum is given by

$$|X(e^{j\omega})|^2 = \prod_{i=1}^{M} \frac{1}{[(\alpha_i^2 + \beta_i^2 - \omega^2)^2 + 4\omega^2\alpha_i^2]} \tag{3.6}$$

and the phase spectrum is given by

$$\theta(\omega) = \angle X(e^{j\omega}) = \sum_{i=1}^{M} tan^{-1}\frac{2\omega\alpha_i}{\alpha_i^2 + \beta_i^2 - \omega^2} \tag{3.7}$$

It is well known that the magnitude spectrum of an individual resonator has a peak at $\omega^2 = \beta_i^2 - \alpha_i^2$ and a half-power bandwidth of α_i. The group delay function can be derived using Equation 3.7 and is given by:

$$\tau(\omega) = \theta^{'}(\omega) = \frac{d\theta(\omega)}{d\omega} = \sum_{i=1}^{M} \frac{2\alpha_i(\alpha_i^2 + \beta_i^2 + \omega^2)}{(\alpha_i^2 + \beta_i^2 - \omega^2)^2 + 4\omega^2\alpha_i^2}. \tag{3.8}$$

For $\beta_i^2 >> \alpha_i^2$, the i^{th} term $\theta_i^{'}(\omega)$, in Equation 3.8, can be approximated around the resonance frequency $\omega_i^2 = \beta_i^2 - \alpha_i^2$, as in [36],

$$\tau(\omega) = \theta_i^{'}(\omega) = \left[\frac{K_i}{(\alpha_i^2 + \beta_i^2 - \omega^2)^2 + 4\omega^2\alpha_i^2} \right] = K_i|H_i(\omega)|^2 \tag{3.9}$$

where K_i is a constant. In this context the group delay function $\tau(\omega)$ has the following properties [36],

- The group delay function $\tau(\omega)$ displays a peak near $\omega^2 = \beta_i^2 - \alpha_i^2$.
- The value of $\tau(\omega)$ is approximately equal to $\frac{2\alpha_i}{\beta_i^2}$, at low frequencies, which is a small constant value.
- The value of $\tau(\omega)$ is approximately equal to $\frac{2\alpha_i}{\omega_i^2}$ at high frequencies.

Hence the group delay function behaves like a squared magnitude response [36].

3.3 RELATIONSHIP BETWEEN SPECTRAL MAGNITUDE AND PHASE

The relation between spectral magnitude and phase has been discussed extensively in [38]. In [38], it has been shown that the unwrapped phase function for a minimum phase signal is given by

$$\theta(\omega) = \theta_v(\omega) + 2\pi\lambda(\omega) = -\sum_{n=1}^{\infty} c(n) sin(n\omega) \tag{3.10}$$

where $c(n)$ are the cepstral co-efficients. Differentiating 3.10 with respect to ω, we have

$$\tau(\omega) = -\theta^{'}(\omega) = \sum_{n=1}^{\infty} nc(n) cos(n\omega) \tag{3.11}$$

where $\tau(\omega)$ is the group delay function. The log-magnitude spectrum for a minimum phase signal $v(n)$ [38], is given by

$$ln|V(\omega)| = \frac{c(0)}{2} + \sum_{n=1}^{\infty} c(n) cos(n\omega) \tag{3.12}$$

The relation between spectral magnitude and phase for a minimum phase signal [38], through cepstral co-efficients, is given by Equations 3.10 and 3.12. For a maximum phase signal Equation 3.10 holds, while the unwrapped phase is given by

$$\theta(\omega) = \theta_v(\omega) + 2\pi\lambda(\omega) = \sum_{n=1}^{\infty} c(n) sin(n\omega) \tag{3.13}$$

and the group delay function $\tau(\omega)$ is given by

$$\tau(\omega) = -\theta^{'}(\omega) = -\sum_{n=1}^{\infty} nc(n) cos(n\omega) \tag{3.14}$$

Hence the relation between spectral magnitude and phase for a maximum phase signal [38], through cepstral co-efficients, is given by Equations 3.10 and 3.13. For mixed phase signals the relation between spectral magnitude and phase is given by two sets of cepstral co-efficients $\{c_1(n)\}$ and $\{c_2(n)\}$, as

$$ln|X(\omega)| = \frac{c_1(0)}{2} + \sum_{n=1}^{\infty} c_1(n) cos(n\omega) \tag{3.15}$$

where $ln|X(\omega)|$ is the log-magnitude spectrum for a mixed phase signal and $\{c_1(n)\}$ is the set of cepstral co-efficients computed from the minimum phase equivalent signal derived from the spectral magnitude. Similarly the unwrapped phase is given by

$$\theta_x(\omega) + 2\pi\lambda(\omega) = -\sum_{n=1}^{\infty} c_2(n) sin(n\omega) \tag{3.16}$$

where $\theta_x(\omega) + 2\pi\lambda(\omega)$ is the unwrapped phase spectrum for a mixed phase signal and $\{c_2(n)\}$ is the set of cepstral co-efficients computed from the minimum phase equivalent signal derived from the spectral phase. Therefore the relation between spectral magnitude and phase for a mixed phase signal [38], through cepstral co-efficients, is given by Equations 3.15 and 3.16.

3.4 THE RESTRICTION OF MINIMUM PHASE

The group delay function can be effectively used for various speech processing tasks only when the signal under consideration is a minimum phase signal. A signal $x_T(n)$

is defined as a minimum phase signal if both $x_T(n)$ and its inverse $x_T^i(n)$ are energy bounded and one sided signals. Alternately as viewed in the z-domain, $x(n)$ is a minimum phase signal if and only if all the poles and zeroes of the z-transform of $x(n)$ lie within the unit circle. Mathematically a minimum phase system is defined as

$$X(z) = \frac{b_0.\Pi_{i=1}^{m}(1 - b_i z^{-1})}{a_0.\Pi_{i=1}^{n}(1 - a_i z^{-1})} \tag{3.17}$$

where $\forall i \; [(b_i < 1) \;\wedge\; (a_i < 1)]$
and $X(z).X^{-1}(z) = 1$.
The phase and magnitude spectra of a minimum phase signal are related through the Hilbert transform [2]. An in-depth analysis of the group delay functions and their relevance for different types of systems like minimum phase, maximum phase etc., can be found in [38].

3.5 PROPERTIES OF GROUP DELAY FUNCTIONS

The group delay functions and their properties have been discussed extensively in [12] and [14]. The two main properties of the group delay functions [14], relevant to this work are

- Additive property.
- High resolution property.

3.5.1 Additive Property

The group delay function exhibits an additive property. If there are two resonators $H_1(e^{j\omega})$ and $H_2(e^{j\omega})$ in a system, then the combined system response is the product of the two individual resonances. Hence the overall response of such a system is given by

$$H(e^{j\omega}) = H_1(e^{j\omega}).H_2(e^{j\omega}) \tag{3.18}$$

where $H_1(e^{j\omega})$ and $H_2(e^{j\omega})$ are the responses of the two resonators whose product gives the overall system response $H(e^{j\omega})$. Taking absolute value on both sides we have

$$|H(e^{j\omega})| = |H_1(e^{j\omega})|.|H_2(e^{j\omega})|, \tag{3.19}$$

Using the additive property of the Fourier transform phase

$$arg(H(e^{j\omega})) = arg(H_1(e^{j\omega})) + arg(H_2(e^{j\omega})). \tag{3.20}$$

Then the group delay function, is given by

$$\begin{aligned} \tau_h(e^{j\omega}) &= -\partial(arg(H(e^{j\omega})))/\partial\omega \\ &= -\partial(arg(H_1(e^{j\omega})))/\partial\omega - \partial(arg(H_2(e^{j\omega})))/\partial\omega \\ \tau_h(e^{j\omega}) &= \tau_{h1}(e^{j\omega})+\tau_{h2}(e^{j\omega}), \end{aligned} \tag{3.21}$$

where, $\tau_{h1}(e^{j\omega})$ and $\tau_{h2}(e^{j\omega})$ correspond to the group delay function of $H_1(e^{j\omega})$ and $H_2(e^{j\omega})$ respectively. From Equations 3.18 and 3.21, it is clear that multiplication in the spectral domain becomes an addition in the group delay domain.

3.5.2 High Resolution Property

The group delay function has a higher resolving power when compared to the magnitude spectrum. The ability of the group delay function to resolve closely spaced formants in the speech spectrum has been investigated in [14]. An illustration is given in Figure 3.1 to highlight the high resolution property of the group delay function over both the magnitude and linear prediction spectrum. In Figure 3.1 (a) is shown the *z*-plane plot of the system consisting of three complex conjugate pole pairs. Figure 3.1 (b) is the corresponding magnitude spectrum, while Figure 3.1 (c) illustrates the spectrum derived using LPC analysis and Figure 3.1 (d) is the corresponding group delay spectrum. One can clearly observe that the three formants are resolved better in the group delay spectrum as in Figure 3.1 (d), when compared to the magnitude spectrum (See Figure 3.1 (b)) or the linear prediction spectrum as in Figure 3.1 (c). The high resolution property is a manifestation of the additive property of the group delay function. To further illustrate the usefulness of the additive property of the group delay spectrum, which helps in resolving two closely spaced resonances (poles), let us consider Figure 3.2, where three different systems are chosen, (i) a complex

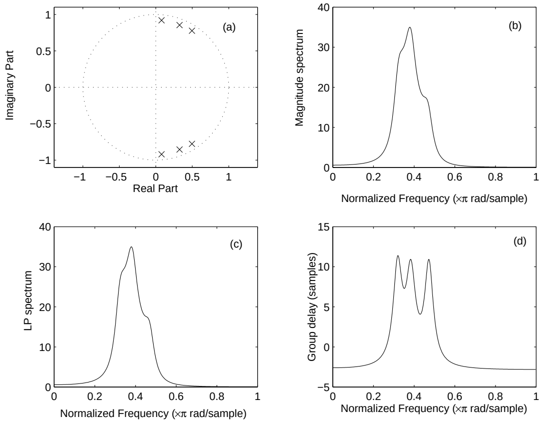

Fig. 3.1: Comparison of the minimum phase group delay function with the magnitude and linear prediction (LP) spectrum. (a) The z-plane with three poles inside the unit circle, (b) the magnitude spectrum of the system shown in (a), (c) the LPC spectrum of the system shown in (a), (d) the group delay spectrum of the system shown in (a).

conjugate pole pair at an angular frequency ω_1, (ii) a complex conjugate pole pair at an angular frequency ω_2 and (iii) two complex conjugate pole pairs one at ω_1, and, the other at ω_2. From the magnitude spectra of these three systems (Figures 3.2(b), 3.2(e) and 3.2(h)), it is observed that even though the peaks in Figure 3.2(b) and Figure 3.2(e) are resolved well, in a system consisting of these two poles, the peaks are not resolved well (see Figure 3.2(h)). This is due to the multiplicative property of magnitude spectra. From Figure 3.2(c), Figure 3.2(f) and Figure 3.2(i), it is evident that in the group delay spectrum obtained by combining the poles together (additive property), the peaks are well resolved as shown in Figure 3.2(i).

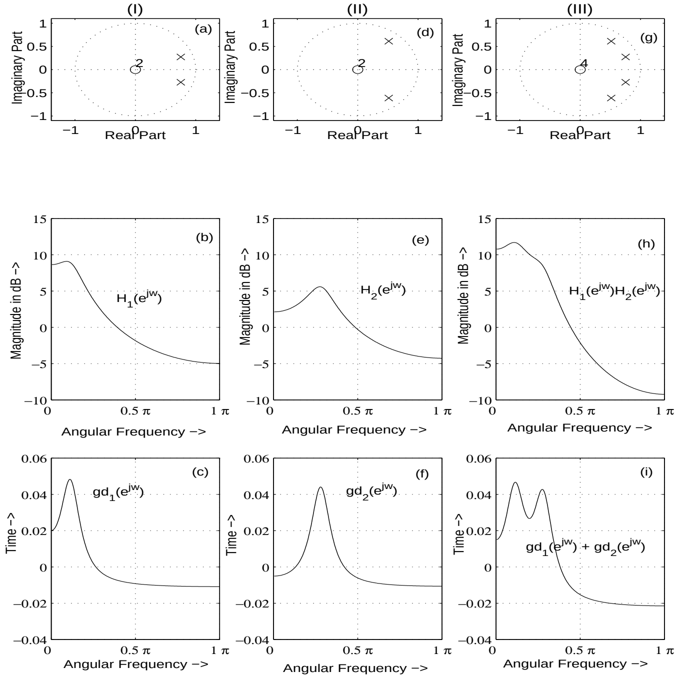

Fig. 3.2: Illustration of the manifestation of the additive property of the group delay function in resolving two closely spaced formants: z-plane, magnitude spectrum and group delay spectrum I) a pole inside the unit circle at $(0.8, \pi/8)$, II) a pole inside the unit circle at $(0.8, \pi/4)$ and III) a pole at $(0.8, \pi/8)$ and another pole at $(0.8, \pi/4)$, inside the unit circle.

3.6 FORMANT ANALYSIS USING THE GROUP DELAY FUNCTION

In this Section, we analyze the formant structure of a synthetic vowel using the group delay function and also discuss related issues. Typically a vowel is characterized by the first three formants. Assuming a source system model of speech production, the transfer function of such a system is given by

$$H(z) = \frac{\sum_{\forall k} b_k z^{-k}}{\sum_{\forall k} a_k z^{-k}} \tag{3.22}$$

The transfer function of the same system for the production of a vowel assuming an all pole model is given by

$$H(z) = \frac{1}{\sum_{k=0}^{k=q} a_k z^{-k}} \tag{3.23}$$

$$H(z) = \frac{1}{1 + \sum_{k=1}^{k=q} a_k z^{-k}} \tag{3.24}$$

assuming $a_0 = 1$ and q is the order of the denominator polynomial.

Let the vowel be characterized by the frequencies F1, F2, F3. Hence the poles of the system are located at

$$p_i = r_i e^{\pm j\omega_i T} \tag{3.25}$$

By substituting Equation 3.25, in Equation 3.23 the system function in equation 3.23 now becomes

$$H(z) = \frac{1}{1 - 2rcos(\omega_i T)z^{-1} + r^2 z^{-2}} \tag{3.26}$$

But from resonance theory

$$r_i = e^{-\pi B_i T} \tag{3.27}$$

By substituting Equation 3.27, in Equation 3.26 the system function in equation 3.26 now becomes :

$$H(z) = \frac{1}{1 - 2e^{-\pi B_i T} cos(\omega_i T)z^{-1} + e^{-2\pi B_i T} z^{-2}} \tag{3.28}$$

In the above array of equations ω is the frequency of the i^{th} formant in radians, B_i is the bandwidth of the i^{th} formant, and T is the sampling rate. Using equation 3.28, a synthetic vowel with two formants at values of F1 = 875 Hz, and F2 = 2300 Hz, B_i = 10% of F_i, and T = 0.0001 s corresponding to a sampling rate of 10 KHz is generated by exciting the system with an impulse train.. The synthetic signal hence generated is used for formant analysis using the group delay function. It is emphasized here that the synthetic vowel includes several pitch periods. Figures 3.3

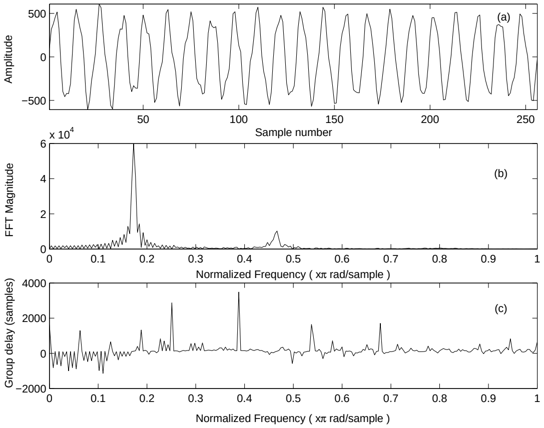

Fig. 3.3: Comparison of the group delay spectrum with the FFT magnitude spectrum for a synthetic vowel. (a) The synthetic vowel with two resonant frequencies, (b) the FFT magnitude spectrum of the signal shown in (a), and (c) the group delay spectrum of the signal shown in (a).

(a), (b), and (c) illustrate the synthetic vowel with two predominant frequencies which are generally called formants in speech, its corresponding magnitude and group delay spectra respectively. The predominant frequencies or the formants are clearly visible in the magnitude spectrum while the group delay function does not display any formant structure. This is primarily because the group delay function for the synthetic vowel becomes spiky due to zeros that are very close to the unit circle in the z-plane and also due to pitch periodicity effects (pitch zeros). Clearly what is required is a modification to the group delay function, that will yield spectra similar to that of the magnitude spectrum. This is what is attempted in the modified group delay function discussed in the next Chapter.

3.7 SUMMARY

This Chapter discusses the theory and properties of group delay functions. The definition of the group delay function and its relation to the magnitude spectrum is described. The definition of a minimum phase signal and its significance in group delay processing is discussed next. The additive and the high resolution properties of the group delay functions are also discussed. The inability of the group delay function to capture formant information is illustrated, which forms the basis for modifying the group delay function. This Chapter discusses the prerequisites for Chapter 4, where the definition of the group delay function is modified to formulate the new modified group delay spectrum.

CHAPTER 4

THE MODIFIED GROUP DELAY FEATURE

The group delay functions can be used to accurately represent signal information as long as the roots of the z-transform of the signal are not too close to the unit circle in the z-plane [39]. The vocal tract system and the excitation contribute to the envelope and the fine structure respectively of the speech spectrum. When the Fourier transform magnitude spectrum is used to extract the formants, the focus is on capturing the spectral envelope of the spectrum and not the fine structure. Similarly the fine structure has to be de-emphasized when extracting the vocal tract characteristics from the group delay function. The zeros that are close to the unit circle manifest as spikes in the group delay function and the strength of these spikes is proportional to the proximity of these zeros to the unit circle. The group delay function becomes spiky in nature also owing to pitch periodicity effects. The spikes introduced by zeros close to the unit circle in the z-plane and also due to pitch periodicity effects, form a significant part of the fine structure and cannot be eliminated by normal smoothing techniques. Hence the group delay function has to be modified to eliminate the effects of these spikes.

4.1 SIGNIFICANCE OF ZERO PROXIMITY TO THE UNIT CIRCLE

The zeros that are close to the unit circle in the z-plane manifest as spikes in the group delay function and the amplitude of these spikes is proportional to the proximity of zeros to the unit circle. To illustrate this a four formant system (with formants at 500 Hz, 1000 Hz, 1500 Hz, and 2500 Hz) characterized by four poles and their complex conjugates is simulated. The pole-zero plot of the four formant system is shown in Figure 4.1 (a), while the corresponding group delay spectrum is shown in Figure 4.1 (b). In Figure 4.1 (c) is shown the pole-zero plot of the same system with zeros added

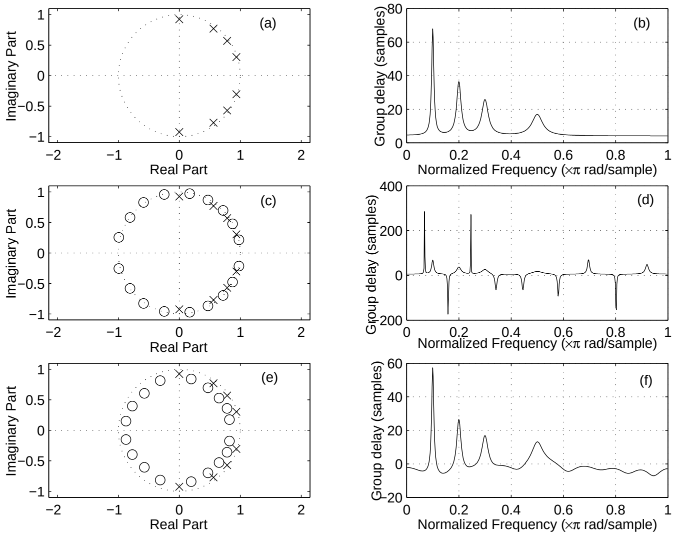

Fig. 4.1: Significance of proximity of zeros to the unit circle (a) The z-plane with four poles inside the unit circle, (b) the group delay spectrum of the system shown in (a), (c) The z-plane with four poles inside the unit circle and zeros added uniformly on the unit circle (d) the group delay spectrum of the system shown in (c), (e) The z-plane with zeros pushed radially inward into the unit circle, (f) the group delay spectrum of the system shown in (e).

uniformly in very close proximity to the unit circle. It is evident from Figure 4.1 (d), that the group delay spectrum for such a system becomes very spiky and ill defined. In Figure 4.1 (e), all the zeros are manually moved (radially) into the unit circle and the group delay function of such a system recomputed. The group delay spectrum of the system in Figure 4.1 (e) is shown in Figure 4.1 (f). It is clear that this technique of pushing the zeros into the unit circle radially, restores the group delay spectrum without any distortions in the original formant locations. This conjecture of restoring the formant structure of the group delay spectrum by pushing the zeros radially inward is utilized in Section 4.3, for modifying the group delay function.

4.2 SIGNIFICANCE OF PITCH PERIODICITY EFFECTS

When the short-time Fourier transform power spectrum is used to extract the formants, the focus is on capturing the spectral envelope of the spectrum and not the fine structure. Similarly the fine structure has to be de-emphasized when extracting the vocal tract characteristics from the group delay function. But the group delay function becomes very spiky in nature due to pitch periodicity effects. This is primarily because the group delay function corresponding to the source completely masks the information about the formants [8]. The zeros that are generated by the impulse train and the finite window lie on the unit circle in the z-plane. Since the group delay function is obtained by sampling the z-transform on the unit circle, the overall group delay function becomes spiky, due to zeros (pitch zeros) very close to the unit circle in the z-plane. To illustrate this a three formant system is simulated whose pole-zero plot is shown in Figure 4.2 (a). The formant locations are at 500 Hz, 1570 Hz, and 2240 Hz. The corresponding impulse response of the system is shown in Figure 4.2 (b) and its group delay function is shown in Figure 4.2 (c). The group delay function is able to resolve all the three formants. The system shown in Figure 4.2 (a) is now excited with 5 impulses and the system response is shown in Figure 4.2 (d). The group delay function of the signal in Figure 4.2 (d) is shown in Figure 4.2 (e). It is evident from Figure 4.2 (e) that the group delay function becomes spiky and distorted due to pitch periodicity effects. The spikes introduced into the group delay function due to zeros close to the unit circle and also due to the pitch periodicity effects, form a significant part of the fine structure and cannot be removed by normal smoothing techniques. Hence the group delay function has to be modified to suppress the effects of these spikes. The considerations discussed in Sections 4.1 and 4.2, form the basis for modifying the group delay function.

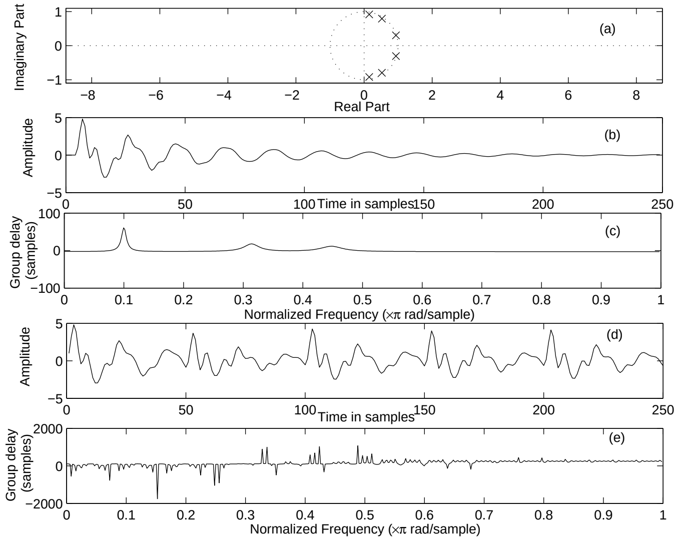

Fig. 4.2: Significance of pitch periodicity effects on the group delay function (a) The z-plane with three complex poles and their complex conjugate pairs inside the unit circle, (b) The impulse response of the system shown in (a), (c) The group delay spectrum of the signal shown in (b), (d) The response of the system shown in (a) to 5 impulses, and (e) The group delay spectrum of the signal shown in (d).

4.3 THE MODIFIED GROUP DELAY FUNCTION

As mentioned in the previous Section 4.1, for the group delay function to be a meaningful representation, it is only necessary that the roots of the transfer function are not too close to the unit circle in the z-plane. Normally, in the context of speech, the poles of the transfer function are well within the unit circle. The zeros of the slowly varying envelope of speech correspond to that of nasals. The zeros in speech are either within or outside the unit circle since the zeros also have non zero bandwidth. In this Section, the computation of the group delay function is modified to to suppress these effects.

A similar approach was taken in [13] for spectrum estimation. Let us reconsider the group delay function derived directly from the speech signal.

$$\tau_x(\omega) = -Im\frac{d(log(X(\omega)))}{d\omega} \tag{4.1}$$

$$= \frac{X_R(\omega)Y_R(\omega) + Y_I(\omega)X_I(\omega)}{|X(\omega)|^2} \tag{4.2}$$

where the subscripts R and I denote the real and imaginary parts of the Fourier transform. $X(\omega)$ and $Y(\omega)$ are the Fourier transforms of $x(n)$ and $nx(n)$, respectively. It is important to note that the denominator term $|X(\omega)|^2$ in Equation 4.2, becomes zero, at zeros that are located close to the unit circle. The spiky nature of the group delay spectrum can be overcome by replacing the term $|X(\omega)|$ in the denominator of the group delay function as in Equation 4.2, with its cepstrally smoothed version, $S(\omega)$.

4.3.1 Significance of Cepstral Smoothing

Assuming a source system model of speech production the z-transform of the system generating the speech signal is given by

$$H(z) = \frac{N(z)}{D(z)} \tag{4.3}$$

where the polynomial $N(z)$ is the contribution due to zeros and the polynomial $D(z)$ is the contribution due to the poles of the vocal tract system. The frequency response of $H(z)$ is given by

$$H(\omega) = \frac{N(\omega)}{D(\omega)} \tag{4.4}$$

where $N(\omega)$ and $D(\omega)$ are obtained by evaluating the polynomials on the unit circle in z-domain. By using the additive property of the group delay function, the group delay function of the system characterized by $H(\omega)$ is given by

$$\tau_h(\omega) = \tau_N(\omega) - \tau_D(\omega) \tag{4.5}$$

where $\tau_N(\omega)$ and $\tau_D(\omega)$ are the group delay functions of $N(\omega)$ and $D(\omega)$ respectively. Spikes of large amplitude are introduced into $\tau_N(\omega)$ primarily due to zeros of $N(z)$

close to the unit circle. As already discussed the group delay function can be directly computed from the speech signal as

$$\tau_x(\omega) = \frac{X_R(\omega)Y_R(\omega) + Y_I(\omega)X_I(\omega)}{|X(\omega)|^2} \tag{4.6}$$

The group delay function for $\tau_N(\omega)$ in Equation 4.5 can be written as

$$\tau_N(\omega) = \frac{\alpha_N(\omega)}{|N(\omega)|^2} \tag{4.7}$$

where $\alpha_N(\omega)$ is the numerator term of Equation 4.6, for $\tau_N(\omega)$. As $|N(\omega)|^2$ tends to zero (for zeros on the unit circle), $\tau_N(\omega)$ has large amplitude spikes. Similarly the group delay function for $\tau_D(\omega)$ in Equation 4.5 can be written as

$$\tau_D(\omega) = \frac{\alpha_D(\omega)}{|D(\omega)|^2} \tag{4.8}$$

where $\alpha_D(\omega)$ is the denominator term of Equation 4.6, for $\tau_D(\omega)$. The term $|D(\omega)|^2$ does not take values very close to zero since $D(z)$ has all roots well within the unit circle. Therefore the term $\tau_D(\omega)$ contains the information about the poles of the system and has no spikes of large amplitude. Substituting equations 4.7 and 4.8 in Equation 4.5, we have :

$$\tau_x(\omega) = \frac{\alpha_N(\omega)}{|N(\omega)|^2} - \frac{\alpha_D(\omega)}{|D(\omega)|^2} \tag{4.9}$$

where $\alpha_N(\omega)$ and $\alpha_D(\omega)$ are the numerator terms of Equation 4.6 for $\tau_N(\omega)$ and $\tau_D(\omega)$ respectively. Assuming that the envelope of $|N(\omega)|^2$ is nearly flat (zero spectrum), multiplying $\tau_x(\omega)$ with $|N(\omega)|^2$ will emphasize the resonant peaks of the second term

$$\frac{\alpha_D(\omega)}{|D(\omega)|^2} \tag{4.10}$$

This leads to the initial form of the modified group delay function which is given by :

$$\tau_m(\omega) = \tau_x(\omega)|N(\omega)|^2 \tag{4.11}$$

Substituting Equation 4.9 in Equation 4.11

$$\tau_x(\omega) = \alpha_N(\omega) - \frac{\alpha_D(\omega)}{|D(\omega)|^2}|N(\omega)|^2 \tag{4.12}$$

In Equation 4.12 an approximation to $|N(\omega)|^2$ is required, which is a nearly flat spectrum (ideally a zero spectrum). An approximation $E(\omega)$ to $|N(\omega)|^2$ can be computed as

$$E(\omega) = \frac{S(\omega)}{S_c(\omega)} \tag{4.13}$$

where $S(\omega)$ is the squared magnitude ($|X(\omega)|^2$)of the signal $x(n)$ and $S_c(\omega)$ is the cepstrally smoothed spectrum of $S(\omega)$ [13, 40]. Alternately the the modified group delay function can be defined as :

$$\tau_m(\omega) = \frac{X_R(\omega)Y_R(\omega) + Y_I(\omega)X_I(\omega)}{S_c(\omega)} \tag{4.14}$$

Therefore the modified group delay function is capable of pushing zeros on the unit circle, radially into the unit circle, and thus emphasizing $\tau_D(\omega)$ which corresponds to the contribution from the poles of the vocal tract system.

4.3.2 Definition of the Modified Group Delay Function

Since the peaks at the formant locations are very spiky in nature, two new parameters γ and α are introduced to reduce the amplitude of these spikes and to restore the dynamic range of the speech spectrum. The new modified group delay function is defined as

$$\tau_m(\omega) = (\frac{\tau(\omega)}{|\tau(\omega)|})\ (|\tau(\omega)|)^{\alpha} \tag{4.15}$$

where

$$\tau(\omega) = (\frac{X_R(\omega)Y_R(\omega) + Y_I(\omega)X_I(\omega)}{S(\omega)^{2\gamma}}) \tag{4.16}$$

where $S(\omega)$ is the smoothed version of $|X(\omega)|$. The two parameters α and γ can vary such that ($0< \alpha \leq 1.0$) and ($0< \gamma \leq 1.0$). Figure 4.3 (a) shows a synthetic signal with two resonances. In Figure 4.3 (b) and (c) are shown the log magnitude and the root magnitude spectrum (root = 2/3) respectively. The group delay and the modified group delay spectrum are shown in Figures 4.3 (d) and (e) respectively. The resonant frequencies are clearly visible in the log magnitude and root magnitude spectrum while the group delay function does not show any structure. This is primarily because, the synthetic signal is non minimum phase. Clearly what is required is a modification to the group delay function, that will yield spectra similar to that of the minimum phase group delay function [14]. This is what is achieved in the modified group delay function as illustrated in Figure 4.3 (e).

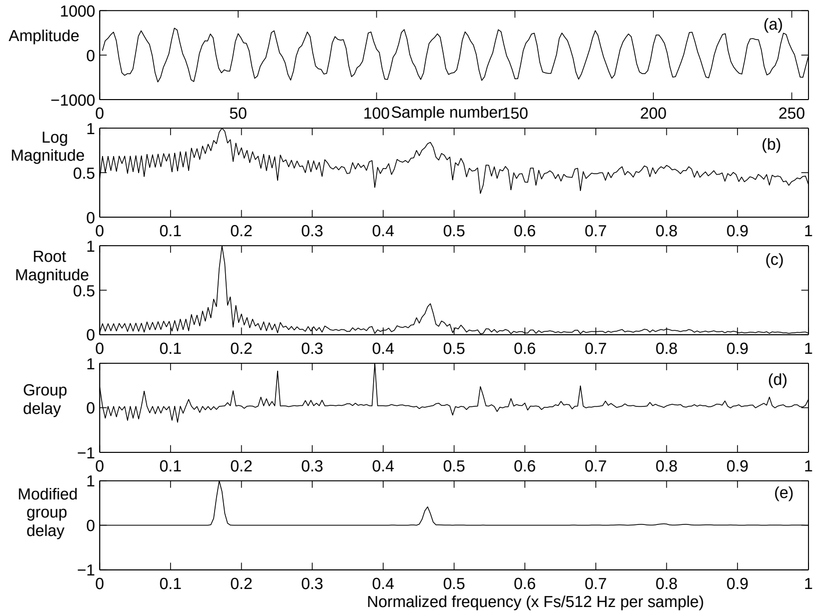

Fig. 4.3: Comparison of various spectra for a synthetic signal (a) The synthetic signal with two resonances (b) the log magnitude spectrum of the signal shown in (a), (c) the root magnitude spectrum (root = 2/3) of the signal shown in (a), (d) the group delay spectrum of the signal shown in (a) and (e) the modified group delay spectrum of the signal shown in (a).

4.4 PARAMETERIZING THE MODIFIED GROUP DELAY FUNCTIO

Since the modified group delay function exhibits a squared magnitude behaviour at the location of the roots the modified group delay function is referred to as the modified group delay spectrum henceforth. Homomorphic processing is the most commonly used approach to convert spectra derived from the speech signal to meaningful features. This is primarily because this approach yields features that are linearly decorrelated which allows the use of diagonal covariances in modeling the speech vector distribution. In this context the discrete cosine transform (DCT I,II,III) [41], is the most commonly used transformation that can be used to convert the modified group delay spectra to cepstral features. Hence the group delay function is converted to cepstra using the discrete cosine transform (DCT II) as

$$c(n) = \sum_{k=0}^{k=N_f} \tau_m(k) \cos(n(2k+1)\pi/N_f) \tag{4.17}$$

where N_f is the DFT order and $\tau_m(k)$ is the modified group delay spectrum. The discrete cosine transform can also be used in the reconstruction of the modified group delay spectra from the modified group delay cepstra. Velocity and acceleration parameters for the new group delay function are defined in the cepstral domain, in a manner similar to that of the velocity and acceleration parameters for MFCC.

4.4.1 Importance of c_0

In the form of the modified group delay cepstrum defined in Equation 4.17, the first coefficient corresponding to $nc(n)$, with $n = 0$ is generally ignored (See Equation 3.4). This value corresponds to the average value in the group delay function. Owing to the effects of linear phase due to the window and the location of pitch peaks with respect to the window, it is really not clear how important the value of $c(0)$ is for recognition. Nevertheless, if we ignore the effects of the window and pitch peaks, the group delay must also contain additional information, in terms of the delays in sources corresponding to that of the formants. This will result in an average value different from zero in the group delay domain. So, it might not be appropriate to ignore the first coefficient in the inverse DCT. Therefore the MODGDF is defined in the form, $\frac{c(n)}{n} + c(0)$ with $n = 1, ..., N-1$ rather than $c(n)$. This is primarily because computation of the MODGDF from the modified group delay spectrum essentially yields $n.c(n)$ as in Equation 3.4. The relation in Equation 3.4 is also discussed in [13].

4.4.2 Algorithm for Computing The Modified Group Delay Cepstra

The following is the algorithm for computing the modified group delay cepstra

- Compute the DFT of the speech signal $x(n)$ as $X(k)$ and the time scaled speech signal $nx(n)$ as $Y(k)$.
- Compute the cepstrally smoothed spectra of $|X(k)|$. Let this be $S(k)$. A low order cepstral window ($lifter_w$) that essentially captures the dynamic range of $|X(k)|$ is chosen.
- Compute the modified group delay function as

  $$\tau_m(k) = (\frac{\tau(k)}{|\tau(k)|})\ (|\tau(k)|)^{\alpha} \tag{4.18}$$

  where

  $$\tau(k) = (\frac{X_R(k)Y_R(k) + Y_I(k)X_I(k)}{S(k)^{2\gamma}}) \tag{4.19}$$

- The two parameters α and γ can vary such that ($0 < \alpha \leq 1.0$) and ($0 < \gamma \leq 1.0$).
- The parameters γ and α need to be tuned appropriately for a given environment.
- Compute the modified group delay cepstra as

  $$c(n) = \sum_{k=0}^{k=N_f} \tau_m(k) \cos(n(2k+1)\pi/N_f) \tag{4.20}$$

  where N_f is the DFT order and $\tau_m(k)$ is the modified group delay spectra.
- The modified group delay cepstra is referred to as the modified group delay feature (MODGDF).

4.5 SUMMARY

The group delay function has to be modified to handle zeros of the z-transform of the signal which are very close to the unit circle in the z-plane and also the pitch periodicity effects. This is achieved by modifying the group delay function. This Chapter discusses the formulation of the modified group delay spectrum. The significance of cepstral smoothing in the formulation of the modified group delay spectrum is also discussed. It is also shown that the modified group delay function is able to capture the formant structure of the speech signal. Cepstral features are extracted from the modified group delay function and are called the modified group delay feature (MODGDF). The algorithm for computing the modified group delay feature (MODGDF) is also briefly enumerated in this Chapter.

CHAPTER 5

FEATURE EVALUATION AND ANALYSIS

In this Chapter, the modified group delay feature (MODGDF) is evaluated using several intuitive and mathematical feature evaluation criteria. The significance of these feature evaluation results in the context of speaker, language and speech recognition is also discussed.

5.1 DECORRELATION

If we assume that the correlation matrix of the input data exhibits Toeplitz structure and neglect boundary effects, the discrete cosine transform always gives decorrelated features. As a proof of concept, we compute the measured correlation matrix for the MODGDF and visualize it, to show that the matrix is diagonal like. Hence diagonal covariances can be used to model the feature vector distribution. The complete decorrelation obtained by the MODGDF is shown as a 3-dimensional plot in Figure 5.1. The first two dimensions correspond to the feature components and the third dimension is the correlation coefficient. It is evident from Figure 5.1, that the diagonal elements are the largest in the correlation matrix when compared to the off-diagonal elements.

5.2 ROBUSTNESS

Features that are invariant to noise save additional processing like cepstral mean subtraction, and eliminate sources of distortion. Representation of speech in the group delay domain enhances important features of the envelope of the short-time speech spectrum making it relatively immune to noise when compared to that of the short-time magnitude spectrum.

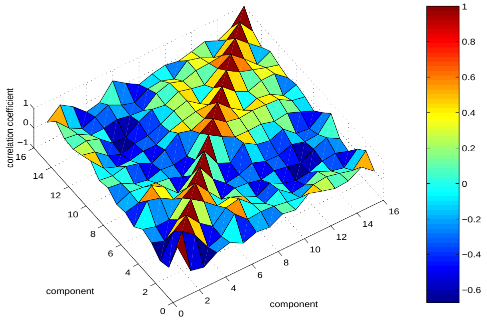

Fig. 5.1: Visualization of the correlation matrix of the MODGD feature set (measured on the NTIMIT database).

5.2.1 Robustness to Convolutional and White Noise

Assuming a source system model of speech production, the clean speech $x_c(n)$, its Fourier transform and the corresponding group delay function [37] is given by

$$x_c(n) = \sum_{k=1}^{p} a_k x_c(n-k) + Ge(n) \tag{5.1}$$

$$X_c(\omega) = \frac{GE(\omega)}{A(\omega)} \tag{5.2}$$

Similarly the noisy speech signal and its Fourier transform are given by:

$$\begin{aligned} x_n(n) &= x_c(n) * h(n) + w(n) && (5.3)\\ X_n(\omega) &= X_c(\omega)H(\omega) + W(\omega) && (5.4) \end{aligned}$$

where h(n) is the time invariant channel response and w(n), the additive white noise. Taking the Fourier transform of Equation 5.1 and substituting in Equation 5.4, $X_n(\omega)$ and the corresponding group delay function $\tau_n(\omega)$ is given by:

$$\begin{aligned} X_n(\omega) &= \frac{GE(\omega)H(\omega) + A(\omega)W(\omega)}{A(\omega)} && (5.5)\\ \tau_n(\omega) &= \tau_{numerator}(\omega) - \tau_a(\omega) && (5.6) \end{aligned}$$

where $\tau_{numerator}(\omega)$ is the group delay function corresponding to that of $GE(\omega)H(\omega)+A(\omega)W(\omega)$ and $\tau_a(\omega)$ is the group delay function corresponding to $A(\omega)$. Further the term $GE(\omega)H(\omega)$ in $\tau_{numerator}(\omega)$ dominates in high SNR regions and the term $A(\omega)W(\omega)$ in $\tau_{numerator}(\omega)$ dominates in low SNR regions. Since α is chosen such that ($0< \alpha \leq 1.0$), the question of noise being emphasized does not arise. In the high SNR case it is the excitation and in the second case it is white noise that makes the group delay spectrum spiky and distorted primarily due to zeros that are very close to the unit circle in the z-domain. White noise has a flat spectral envelope and hence contributes zeros very close to the unit circle. Further the locations and amplitudes of these spikes is also not known. To suppress these spikes, the behavior of the spectrum where the noise zeros contribute to sharp nulls is utilized. A spectrum with a near flat spectral envelope containing the spectral shape contributed by the zeros is derived using cepstral smoothing as discussed in Section 4.3.1 and multiplied with the group delay function to get the modified group delay function as in Equations 4.18 and 4.19. The effects due to the excitation can be dealt with by pushing all zeros very close to the unit circle in the z-domain, well inside the unit circle by appropriately selecting values for the two parameters α and γ as defined in Equations 4.18 and 4.19.

5.2.2 Comparison to Log and Root Compressed Cepstra

In general, log-cepstral analysis is sensitive to noise and the root compressed cepstral approaches [42, 43] represent speech better in noise. In this Section, we compare the the log and root compression approaches with the MODGDF in the presence of white noise at different values of SNR. We pick one complete sentence *"Critical equipment needs proper maintenance"* from the TIMIT database. This sentence is added with white noise scaled by a factor η. The value of η is varied and the SNR computed. The average error distributions between the clean and the noisy speech across all frames corresponding to the same sentence are then calculated for four different values of SNR at 0, 3, 6, and 10 dB. In Figure 5.2, the average error distributions of the the MODGDF ($\alpha = 0.4$, $\gamma = 0.9$), the spectral root compressed cepstra [42] (root =

2/3), the energy root compressed cepstra [43] (root $= 0.08$), and the log compressed cepstra (MFCC) are compared. Figures 5.2 (a), (b), (c), and (d) correspond to the

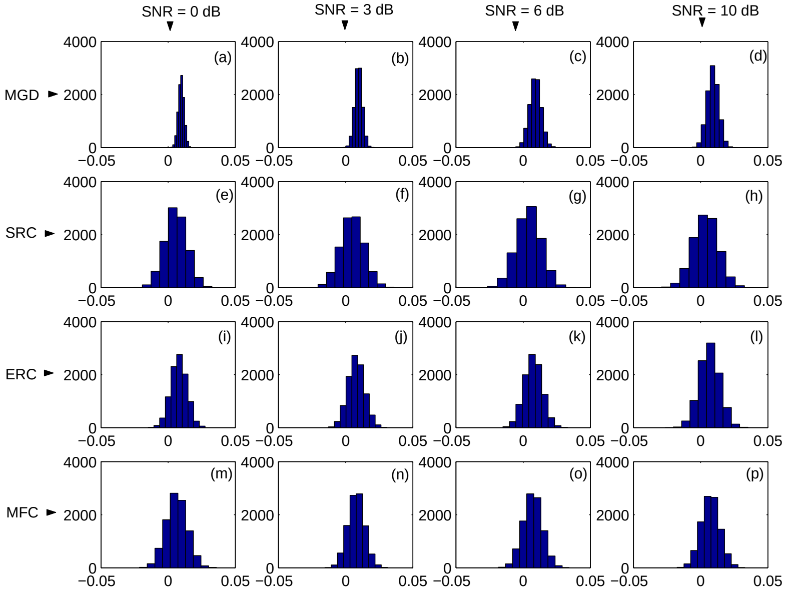

Fig. 5.2: Comparison of the average error distributions (AED) of the MODGDF and root compressed cepstra in noise. (a) AED of the MODGDF ($\alpha = 0.4$, $\gamma = 0.9$) at 0 dB SNR, (b) AED of MODGDF at 3 dB SNR, (c) AED of MODGDF at 6 dB SNR, (d) AED of MODGDF at 10 dB SNR, (e) AED of the spectrally root compressed (SRC) cepstra (root $= 2/3$) at 0 dB SNR, (f) AED of SRC at 3 dB SNR, (g) AED of SRC at 6 dB SNR, (h) AED of SRC at 10 dB SNR, (i) AED of the energy root compressed (ERC) cepstra (root $= 0.08$) at 0 dB SNR, (j) AED of ERC at 3 dB SNR, (k) AED of ERC at 6 dB SNR, (l) AED of ERC at 10 dB SNR, (m) AED of the mel frequency (MFC) cepstra at 0 dB SNR, (n) AED of MFC at 3 dB SNR, (o) AED of MFC at 6 dB SNR, and (p) AED of MFC at 10 dB SNR.

average error distribution of the MODGDF computed for a SNR of 0, 3, 6, and 10 dB respectively, while Figures 5.2 (e), (f), (g), and (h) correspond to the average error distribution of the spectral root compressed cepstra (SRC) computed for a SNR of 0, 3, 6, and 10 dB respectively. In Figures 5.2 (i), (j), (k), and (l) are shown the average error distribution of the the energy root compressed cepstra (ERC) computed for a SNR of 0, 3, 6, and 10 dB respectively, while Figures 5.2 (m), (n), (o), and (p) correspond to the average error distribution of the log compressed cepstra (MFC) computed for a SNR of 0, 3, 6, and 10 dB respectively. It is clear from the Figure 5.2 that average deviation of the noisy speech cepstra from the clean speech cepstra is the least for the MODGDF when compared to either the spectral root, the energy root or the log compressed cepstra.

5.2.3 Similarity to RASTA

RASTA (RelAtive SpecTrA) [44], is a popular technique used to handle speech degraded with both convolutional and white noise. RASTA filters out very low temporal frequency components below 1 Hz which are primarily due to the changing auditory environment. It also filters out higher frequency temporal components greater than 13 Hz as they represent changes faster than the rate at which the speech articulators can move. But the issue on which we propose to compare RASTA with the MODGDF is the use of a compressing static nonlinear transformation (generally a logarithm operation) on the critical band spectrum. Instead of compressing the PLP spectrum logarithmically as in RASTA, the group delay spectrum is raised to the power of α. It is emphasized here that, by varying the value of α a control over the time trajectories of spectral components can be exercised. Further by varying the value of γ, control over the zeros lying on the unit circle can be gained. Hence the MODGDF has one relative advantage over the RASTA technique. It avoids additional processing steps of RASTA, like deriving the critical band spectrum and a compressing non linear transformation.

5.2.4 Significance of Cepstral Mean Subtraction

Cepstral mean subtraction (CMS) is a successful technique used to filter out linear distortions introduced into the speech signal due to the telephone channel. Let T(z) be the *z*-transform of a telephone speech signal

$$T(z) = S(z)G(z) \tag{5.7}$$

where S(z) is the *z*-transform of clean speech and G(z) the *z*-transform of the channel. In the log domain

$$logT(z) = logS(z) + logG(z) \tag{5.8}$$

The cepstrally mean subtracted vector is given by :

$$c_{cms}(n) = c_{mgd}(n) - E[c_{mgd}(n)] \tag{5.9}$$

where $E[c_{mgd}(n)]$ is the expectation of the modified group delay cepstra taken over a number of frames of channel corrupted speech. It is emphasized in [44], that CMS is capable of handling convolutive noise only and therefore RASTA with CMS always significantly improves the performance of a speech recognition system. Applying CMS on the MODGDF also gives an improvement in recognition performance.

5.2.5 Significance of Removing Channel Effects in the Group Delay Domain

Owing to the nonlinearities that are introduced by $\alpha \neq 1$ and $\gamma \neq 1$, the removal of the channel effects is an issue. It is clear therefore that, if the channel effects are multiplicative, they become additive in the phase and hence the group delay domain, provided that γ and α are each equal to 1. Generally, in the case of MFCC, the mean removal is done in the cepstral domain. In the case of the modified group delay function, owing to the artifacts introduced by α and γ it is not clear in which domain, the mean removal must be performed. Hence two different approaches have been tried

- Ignore the cross terms, assume that channel effects are additive in the cepstral domain.

- Perform noise removal in the group delay domain with α and γ set to one.

Although the second approach is theoretically correct, the performance of the system as noticed in our experiments using the first approach seems to be far superior. This could be due to the fact, that the signal is not only corrupted by multiplicative channel effects but also additive noise. Using the argument in [13], it is important to suppress the effects of noise in the modified group delay function before it can be further processed. In the context of the second approach, the other issue would be whether the mean removal should be performed on the envelope of the modified group delay function or on the standard modified group delay function. This is similar to converting the raw Fourier spectrum into filter bank energies in the computation of MFCC. To enable this, a new parameter $lifter_w$ (See Section 6.2.1) is introduced. This parameter defines the fineness of the envelope of the modified group delay function for mean computation.

5.3 ANALYSIS OF TASK INDEPENDENCE USING THE SEQUENTIAL FORWARD SEARCH

Feature selection techniques have been widely used in pattern recognition primarily to prune features that do not contribute to discrimination among classes. Although a multitude of feature selection techniques exist in pattern recognition literature, only a few of them suit speech recognition applications, where recognition rate must be used as an internal criterion for selecting features. The sequential forward search (SFS) is one such technique. In this Section, we use the SFS technique [45], to evaluate the the MODGD features using the Bhattacharya distance metric [45]. The Bhattacharya distance measure is used to investigate class separability criteria in this context since the MODGDF shows good separability on a pairwise basis as in [46]. Indeed the Bhattacharya distance is primarily defined for a two class problem and extension to a multi-class case is a combination of pairwise bounds as illustrated in the Equations 5.10 and 5.11. The Bhattacharya distance for a two class and a multi (M) class case

is given by [45]

$$B_{pair}(X) = \frac{1}{2}\int [p(X|\omega_1)p(X|\omega_2)]^{1/2} dX \tag{5.10}$$

$$B_M(X) = \frac{1}{2}\sum_{i>j}^{M}\sum_{j=1}^{M}\int [p(X|\omega_i)p(X|\omega_j)]^{1/2} dX \tag{5.11}$$

where Equation 5.10 gives the Bhattacharya distance between two class and Equation 5.11 gives the Bhattacharya distance between M classes. We use the following algorithm for the calculation of the Bhattacharya distance using sequential forward search (SFS) [45]:

```
1. Start with an empty set $\mathcal{P} = \{\phi\}$, as the current set of selected features
2. Let $\mathcal{Q}$ be the full set of 16 dimensional MODGD features
3. While the size of $\mathcal{P}$ is less than 16
    (a) for each $\upsilon \in \mathcal{Q}$
        i. set $\mathcal{P}' \leftarrow \{\upsilon\} \cup \mathcal{P}$
        ii. compute the Bhattacharya distance $\upsilon^*$ with $\mathcal{P}'$
    (b) set $\mathcal{P}' \leftarrow \{\upsilon^*\} \cup \mathcal{P}$
    (c) set $\mathcal{Q} \leftarrow$ Q$\setminus$ $\{\upsilon^*\}$
    (d) save the Bhattacharya distance calculated with the current $\mathcal{P}$
4. Return the MODGD feature number and the corresponding separability criterion (Bhattacharya distance)
```

The histograms corresponding to the separability criterion (Bhattacharya distance) versus the feature dimension, calculated using Equation 5.11 are shown in Figure 5.3 (a), for speaker separability, and in Figure 5.3 (b), for language separability. The separability criteria are calculated for a 50 speaker (NTIMIT data [4]), and a 11 language (OGI data [47]), task respectively. It is emphasized here that the MODGDF is computed with $\alpha = 0.4$ and $\gamma = 0.9$ in both these cases. It is worthwhile to note that

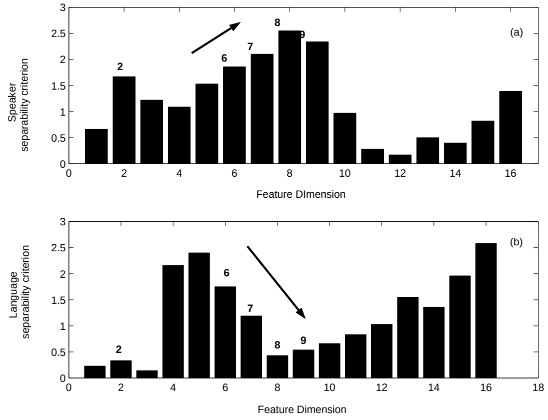

Fig. 5.3: Histograms of Bhattacharya distance criteria versus feature dimension for (a) Speaker separability, and (b) Language separability.

dimensions 2, 6, 7, 8, and 9 of the MODGDF are very useful for discriminating speakers while the same dimensions are least useful for discriminating languages. Further dimensions 6, 7, 8, and 9 fall in the region of rise (indicated by an arrow) for speaker discrimination while the same region corresponds to a region of fall for language discrimination. It is therefore convenient to state that the MODGDF is task independent in the sense that different dimensions capture different information (speaker, language related).

5.4 CLUSTER STRUCTURE AND SEPARABILITY ANALYSIS IN THE 2-DIMENSIONAL FEATURE SPACE

The goal of feature extraction is to find a transformation to a relatively low dimensional feature space that preserves the information pertinent to the recognition problem and to enable acceptable comparisons to be made. In this Section, we use Sammon

mapping [48], which belongs to the class of multi-dimensional scaling techniques, for dimensionality reduction and separability analysis of the MODGDF.

5.4.1 Separability Analysis using Sammon Mapping

Sammon mapping is an iterative method based on gradient search [48]. Sammon mapping minimizes the following error function

$$E_{sam} = \frac{1}{\sum_{i=1}^{i=N-1} \sum_{j=i+1}^{i=N} D_{ij}} \sum_{i=1}^{i=N-1} \sum_{j=i+1}^{i=N} \frac{(d_{ij} - D_{ij})^2}{D_{ij}} \tag{5.12}$$

where d_{ij} is the distance between two points i,j in the d-dimensional output space, and D_{ij} is the distance between two points i , j in the D-dimensional input space, N is the number of points in the input or output space. In this Section, an effort has been made to visualize two dimensional codebooks for a pair of speakers and also a pair of languages using Sammon mapping. Each speaker's and language's codebook of size thirty

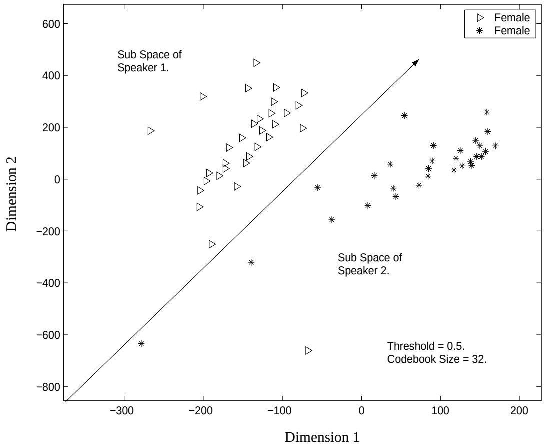

Fig. 5.4: Two dimensional visualization of female-female speaker discrimination with the MODGDF using Sammon mapping.

two is generated by concatenating six sentences of that particular speaker picked from the training set of the NTIMIT [4] database. The codebook which consists of thirty two, sixteen dimensional code vectors is transformed into a two dimensional codebook of size thirty two after Sammon mapping [48]. Figure 5.4, shows the distribution of the code vectors for two female speakers using the MODGDF. It can be observed that in Figure 5.4 the code vectors corresponding to each of the speakers can be separated by a straight line. Similar results are demonstrated for a pair of languages English and French in Figure 5.5. Note that the the code vectors of the two languages English and French can be separated by a straight line. It is emphasized here that as the number of speakers or languages is increased the code vectors tend to overlap in the two dimensional feature space.

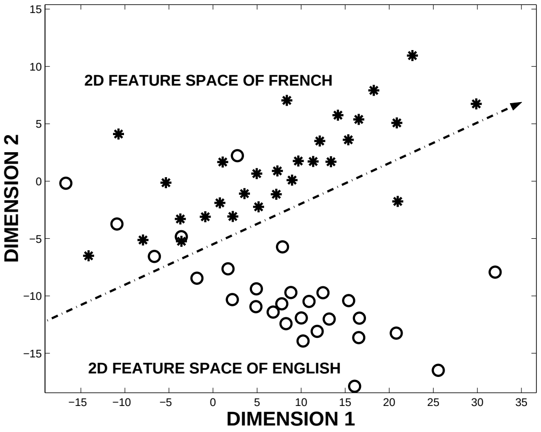

Fig. 5.5: Two dimensional visualization of English-French language discrimination with the MODGDF using Sammon mapping.

5.5 SEPARABILITY ANALYSIS IN THE HIGH-DIMENSIONAL FEATURE SPACE

The most commonly used separability measures in speech recognition are the geometrically intuitive measures like the F-ratio and mathematical measures like the Chernoff and the Bhattacharya bound [45]. The Bhattacharya bound which is a special case of the Chernoff bound is a probabilistic error measure and relates more closely to the likelihood maximization classifiers that we use for performance evaluation. The results of analysis presented herein are for measuring class separability between speakers, in the context of speaker identification, and languages, in the context of language identification. We refer to speakers and languages as classes, following the general practice in pattern recognition terminology, in the analysis that follows. The Bhattacharya distance [45], B is defined as

$$B(X) = -ln \int [p(X|\omega_1)p(X|\omega_2)]^{1/2} dX \tag{5.13}$$

Assuming that the distributions are Gaussian, the probability density function for the i th class is given by

$$p(X|\omega_i) = [(2\pi)^n|\Sigma_i|]^{-\frac{1}{2}} e^{(-1/2)(X-\mu_i)^T\Sigma_i{}^{-1}(X-\mu_i)} \tag{5.14}$$

where μ_i is the mean vector and Σ_i is the covariance matrix of the i th class distribution. The multivariate integral in Equation 5.13 can be evaluated and simplified to

$$B = \frac{1}{8}(\mu_2-\mu_1)^T\left[\frac{\Sigma_1+\Sigma_2}{2}\right]^{-1}(\mu_2-\mu_1) + \frac{1}{2}ln\left[\frac{|1/2(\Sigma_1+\Sigma_2)|}{\sqrt{|\Sigma_1||\Sigma_2|}}\right] \tag{5.15}$$

Assuming that the feature components are independent of each other and from Equation 5.15 the distance between any two feature vectors f_k and f_l can be computed on a component pair basis. We can therefore define the distance D_i between the component pairs i of the two feature vectors f_k and f_l as

$$D_i(f_k, f_l) = \frac{1}{4}\frac{(\mu_{ik}-\mu_{il})^2}{(\sigma_{ik}^2+\sigma_{il}^2)} + \frac{1}{2}ln\left[\frac{1/2(\sigma_{ik}^2+\sigma_{il}^2)}{\sqrt{\sigma_{ik}^2\sigma_{il}^2}}\right] \tag{5.16}$$

Finally the Bhattacharya distance D between the two feature vectors f_k and f_l with n number of component pairs is given by

$$D(f_k, f_l) = \sum_{i=1}^{n} D_i(f_k, f_l) \tag{5.17}$$

Further the Bhattacharya distance for a two class and a multi (M) class case as in [45], is given by Equation 5.10 and Equation 5.11 respectively. The MODGDF shows good separability on a pairwise basis as in Section 5.4.1 and therefore we use the Bhattacharya distance measure to investigate class separability criteria. We therefore consider 50 speakers from the NTIMIT [4] database and compute a 16 dimensional codebook of size 32 for each speaker. Similarly we consider 11 languages from the OGI_MLTS [47] database and compute a 16 dimensional codebook of size 32 for each language. The cumulative separability criterion based on the Bhattacharya distance measure is then calculated. The cumulative speaker separability criterion versus feature dimension for the MODGDF and the MFCC is illustrated in Figure 5.6 (a). The cumulative language separability criterion versus feature dimension for both the MODGDF and the MFCC is illustrated in Figure 5.6 (b). From Figures 5.6 (a) and (b), it is clear that the MODGDF outperforms MFCC with respect to class separability for both the speaker and the language tasks, as the cumulative separability curve corresponding to MODGDF is above that of the MFCC.

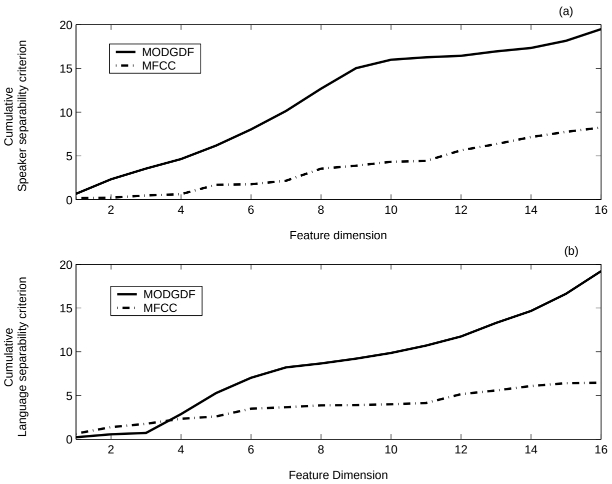

Fig. 5.6: Results of Separability analysis. (a) Cumulative speaker separability of MODGDF and MFCC using Bhattacharya distance, and (b) Cumulative language separability of MODGDF and MFCC using Bhattacharya distance.

5.6 SUMMARY

This Chapter evaluates the modified group delay feature (MODGDF) using several feature evaluation criteria. The decorrelation obtained by the MODGDF is illustrated by visualizing a measured correlation matrix. The MODGDF is analysed for robustness to both convolutional and white noise. Similarity of the MODGDF to RASTA and the significance of cepstral mean subtraction in the the modified group delay domain is discussed. The task independence of the MODGDF is illustrated using the sequential forward search and the Bhattacharya distance metric. The results of cluster structure and separability analysis in the 2-dimensional feature space using Sammon mapping indicates that the MODGDF is able to linearly separate classes on a pairwise basis. It is shown that the cumulative separability of the MODGDF is better than the MFCC when the Bhattacharya distance measure is used.

CHAPTER 6

PERFORMANCE EVALUATION

In this Chapter, the MODGDF is used as a front end for building automatic speaker, language, and syllable recognition systems. The computation of the MODGDF and other features used in the study is discussed first. The procedures adopted to estimate the optimal values for the parameters $lifter_w$ (length of the window used in the cepstral domain), α, and γ, that give the best recognition performance across all the three tasks are also described. The performance of the MODGDF is compared to that of MFCC, which are perhaps the most commonly used features in speech recognition today. Since there are no filter banks involved in the computation of the MODGDF, it's performance is compared with LFCC. In the computation of the MODGDF the modified group delay spectrum is compressed by a root value. Hence it's performance is also compared with spectral root compressed MFCC [42], and energy root compressed MFCC [43]. A formant reconstruction algorithm is proposed to highlight the significance of combining features. The MODGDF is combined with other Fourier transform magnitude-based features both at the feature and the decision level. The combined features are used for all the aforementioned three tasks and the results discussed. Combining the MODGDF and the MFCC gives a significant increase in recognition performance, while combining any two features derived from the short time magnitude spectra like the MFCC and the LFCC does not give any improvement in recognition performance.

6.1 DATABASES USED IN THE STUDY

Since the MODGDF has been used for the tasks of syllable, speaker, language recognition, there are four databases used in the study. The databases used are TIMIT [3] and NTIMIT [4] for speaker identification, OGI_MLTS [47] and the DBIL [5] for language identification, and the Database for Indian languages (DBIL) [5] for syllable recognition.

6.1.1 The TIMIT Database

The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus [3] was a joint effort among the Massachusetts Institute of Technology (MIT), Stanford Research Institute (SRI), and Texas Instruments (TI). TIMIT contains a total of 6300 sentences, 10 sentences spoken by each of 630 speakers, from 8 major dialect regions of the United States.

6.1.2 The NTIMIT Database

The NTIMIT corpus [4], was developed by the NYNEX Science and Technology Speech Communication Group to provide a telephone bandwidth adjunct to the popular TIMIT Acoustic-Phonetic Continuous Speech Corpus. NTIMIT was collected by transmitting all 6300 original TIMIT utterances though various channels in the NYNEX telephone network and re-digitizing them. The actual telephone channels used were varied in a controlled manner, in order to sample various line conditions. The NTIMIT utterances were time-aligned with the original TIMIT utterances so that the TIMIT time-aligned transcriptions could be used with the NTIMIT corpus as well.

6.1.3 The OGI_MLTS Database

The OGI Multi-language Telephone Speech Corpus [47], consists of telephone speech from 11 languages. The initial collection, included 900 calls, 90 calls each in 10 languages and was collected by Muthusamy [47]. The languages are English, Farsi, French, German, Japanese, Korean, Mandarin, Spanish, Tamil and Vietnamese. It is from this initial set that the training (50), development (20) and test (20) sets were established. The National Institute of Standards and Technology (NIST) uses the same 50 - 20 - 20 set that was established. The corpus is used by NIST for evaluating the performance of automatic language identification systems.

6.1.4 The Database for Indian Languages (DBIL)

The Database for Indian languages (DBIL) [5], collected by Speech and Vision Lab at IIT Madras, India, consists of broadcast news bulletins of two Indian languages, Tamil and Telugu.

- DBIL Tamil database [5]: This corpus consists of 20 news bulletins of Tamil language transmitted by Doordarshan India, each of 15 minutes duration comprising 10 male and 10 female speakers. The total number of distinct syllables is 2184.
- DBIL Telugu database [5]: This corpus consists of 20 news bulletins of Telugu language transmitted by Doordarshan India, each of 15 minutes duration comprising 10 male and 10 female speakers. The total number of distinct syllables is 1896.

6.2 COMPUTATION OF VARIOUS FEATURES

In this Section, the computation of the MODGDF and the other features like MFCC, LFCC, spectral root compressed MFCC and energy root compressed MFCC are discussed.

6.2.1 Algorithm for Computing the Modified Group Delay Cepstra

The following is the algorithm for computing the modified group delay cepstra

- Pre-emphasize the speech signal x(n), followed by frame blocking, at a frame size of 20 ms and frame shift of 10 ms. A hamming window is applied on each frame of the speech signal.
- Compute the DFT of the framed and windowed speech signal $x(n)$, as $X(k)$, and the time scaled speech signal $nx(n)$, as $Y(k)$.
- Compute the cepstrally smoothed spectra of $|X(k)|$. Let this be $S(k)$. A low order cepstral window ($lifter_w$) that essentially captures the dynamic range of $|X(k)|$ should be chosen.

- Compute the modified group delay function as:

$$\tau_m(k) = (\frac{\tau(k)}{|\tau(k)|})\ (|\tau(k)|)^{\alpha} \tag{6.1}$$

where

$$\tau(k) = (\frac{X_R(k)Y_R(k) + Y_I(k)X_I(k)}{S(k)^{2\gamma}}) \tag{6.2}$$

- The two parameters α and γ can vary such that, (0< $\alpha \leq 1.0$) and (0< $\gamma \leq 1.0$).
- Set the value $\alpha = 0.4$ and $\gamma = 0.9$ (See Section 6.3 and 6.3.3 for estimation of these values of α and γ). Other values of α and γ, where (0< $\alpha \leq 1.0$) and (0< $\gamma \leq 1.0$) can also be determined for a particular environment by using line search (See Section 6.3.3).
- Compute the modified group delay cepstra as

$$c(n) = \sum_{k=0}^{k=N_f} \tau_m(k) \cos(n(2k+1)\pi/N_f) \tag{6.3}$$

where N_f is the DFT order and $\tau_m(k)$ is the modified group delay spectra.
- The modified group delay cepstra is referred to as the modified group delay feature (MODGDF).
- The velocity, acceleration and energy parameters are added to the MODGDF in a conventional manner.

6.2.2 Extraction of MFCC

The speech signal is first pre-emphasized and transformed to the frequency domain using a fast Fourier transform (FFT) . The frame size used is 20ms and the frame shift used is 10ms. A hamming window is applied on each frame of speech prior to the computation of the FFT. The frequency scale is then warped using the bilinear transformation proposed by Acero [49].

$$\omega_{warped} = \omega + 2tan^{-1}\frac{F_\omega sin\omega}{1 - F_\omega cos\omega} \tag{6.4}$$

where the constant F_ω, which varies from 0 to 1, controls the amount of warping. The frequency scale is then multiplied by a bank of filters N_f, whose center frequencies are uniformly distributed in the interval $[Min_f, Max_f]$, along the warped frequency axis. Min_f is the minimum frequency and Max_f is the maximum frequency which primarily decide the useful frequency range of the particular speech data being handled. The filter shape used at the front end is trapezoidal and its width varies from one center frequency to another. The shape of the filter is controlled by a constant which varies from 0 to 1, where 0 corresponds to triangular and 1 corresponds to rectangular. The filter bank energies are then computed by integrating the energy in each filter. A discrete cosine transform (DCT) is then used to convert the filter bank log energies to cepstral co-efficients. Cepstral mean subtraction is always applied when working with telephone speech. A perceptually motivated filter design is also used as in [34]. The front end parameters are tuned carefully as in [34] for computing the MFCC so that the best performance is achieved. The LFCC are computed in a similar fashion except that the frequency warping is not done as in the computation of the MFCC. The velocity, acceleration and the energy parameters are added for both the MFCC and LFCC in a conventional manner.

6.2.3 Extraction of Spectral root and Energy Root Compressed MFCC

The spectral root compressed MFCC are computed as described in [42] and the energy root compressed MFCC as in [43]. The computation of the spectral root compressed MFCC is the same as the computation of the MFCC except that instead of taking a log of the FFT spectrum we raise the FFT spectrum to a power γ where the value of γ ranges from 0 to 2. In the computation of the energy root compressed MFCC instead of raising the FFT spectrum to the root value, the Mel frequency filter bank energies are compressed using the root value. In the energy root compressed case the value of the root used for compression can range from 0 to 1. It is emphasized here that the front end parameters involved in the computation of both these features including the root value have been tuned carefully so that they give the best performance and are not handicapped in any way when they are compared with the MODGDF. The value of the spectral root and the energy root used in the experiments are 2/3 and 0.08 respectively. The velocity, acceleration and the energy parameters are augmented to both forms of the root compressed MFCC in a conventional manner.

6.3 ESTIMATION OF OPTIMAL VALUES FOR $LIFTER_W$, α, AND γ

In this Section, the optimal values of the three free parameters $lifter_w$, α, and γ used in the computation of the MODGDF are estimated from a signal processing perspective. We first fix the range of values that these three parameters can take.

- The length of the cepstral window $lifter_w$ can vary from 4 to 9 for capturing the envelope of the speech spectrum.
- The parameter α can vary between 0 and 1.
- The parameter γ can vary between 0 and 1.

We substantiate the above conjectures from a signal processing viewpoint in the following Sections 6.3.1 and 6.3.2.

6.3.1 Estimation of Optimal Values for $lifter_w$

As discussed in Section 4.3.1, the problem of restoring the resonant structure of the signal with the modified group delay function is reduced to the estimation of $|N(\omega)|^2$ by a near flat spectrum $E(\omega)$ (See Equation 4.13). In practice $|N(\omega)|^2$ has to be estimated from the signal. The values of $|N(\omega)|^2$ around the zeros have to be preserved so that they cancel the small values in the denominator of the first term in Equation 4.5. The selection of the length of the window $lifter_w$, used for cepstral smoothing is crucial in terms of capturing the correct formant structure of the speech signal. The value of $lifter_w$ needs to be set at the appropriate value to obtain the best recognition performance. A series of initial experiments conducted for phoneme recognition in [37] showed that any value of $lifter_w$ greater than 9 hurts performance. Hence the length of $lifter_w$ is selected to vary from 4 to 9. From a signal processing perspective, the estimate $E(\omega)$ as in Equation 4.13 should result in a flat spectrum. In this Section, we show that even for smaller lengths (4 to 9) of $lifter_w$, the estimated spectrum $E(\omega)$ is indeed flat, by considering a short segment of speech. The short segment of speech considered for analysis is shown in Figures 6.1 (a) and (b). The squared magnitude spectrum $S(\omega)$ and its cepstrally smoothed version of $S_c(\omega)$ are shown in Figure 6.1 (c) and (d), for a value of $lifter_w = 6$ and $lifter_w = 16$, respectively. In Figure 6.1 (e), is shown the estimated flat spectrum $E(\omega)$, for a value of $lifter_w = 6$. The estimated flat spectrum $E(\omega)$, for a value of $lifter_w = 16$, is shown in Figure 6.1 (f). It is clear from Figure 6.1 (f), that the estimated spectrum is flat for longer window lengths (> 9). From Figure 6.1 (e), it is clear that the estimated spectrum $E(\omega)$ is indeed flat even for shorter window lengths (4 to 9). Hence the the value of the length of the window in the cepstral domain $lifter_w$, can be fixed in the range from 4 to 9.

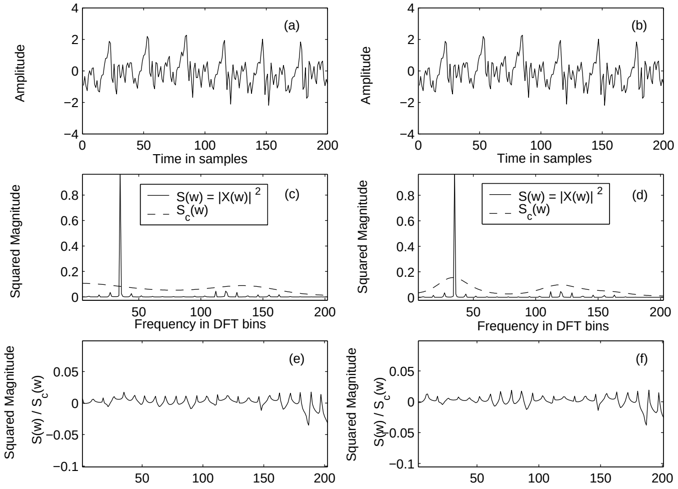

Fig. 6.1: Comparison of the estimated flat spectrum for different cepstral window lengths. (a) & (b) A short segment of speech, (c) The squared magnitude spectrum $S(\omega)$ and its cepstrally smoothed version of $S_c(\omega)$ for a value of $lifter_w = 6$, (d) The squared magnitude spectrum $S(\omega)$ and its cepstrally smoothed version of $S_c(\omega)$ for a value of $lifter_w = 16$, (e) The estimated flat spectrum $E(\omega)$, for a value of $lifter_w = 6$, and (f) The estimated flat spectrum $E(\omega)$, for a value of $lifter_w = 16$.

6.3.2 Estimation of Optimal Values for α and γ

In this analysis, the value of the $lifter_w$ is fixed at 6, although any variation from 4 to 9 has little effect on the envelope of the modified group delay spectra as discussed in Section 6.3.1. The effects of pitch periodicity make the modified group delay function spiky at formant locations [50]. Hence in order to fix the values of α and γ we consider a system characterized by 3 formants (3 complex conjugate pole pairs) as in Figure 6.2 (a). The system in Figure 6.2 (a) is excited with an impulse and the corresponding impulse response is shown in Figure 6.2 (b). From a signal processing perspective this is equivalent to a signal with a single pitch period. The group delay spectrum of the response in 6.2 (b) is shown in6.2 (d). The system in Figure 6.2 (a) is excited with a train of 5 impulses spaced apart by 60 samples and the corresponding impulse response is shown in Figure 6.2 (c). From a signal processing perspective this is equivalent to a signal with five pitch periods. The group delay spectrum of the response in 6.2 (c) is shown in 6.2 (e), and has no structure or formant information. In Figure 6.2 (f), is shown the envelope of modified group delay spectrum for $lifter_w = 6$, $\alpha = 1$, and $\gamma = 1$. It is clear that for these values of $lifter_w$, α, and γ, the envelope of the spectrum of speech is incorrectly captured by the modified group delay spectrum. Hence the values of α, and γ need to be fixed such that the formant locations are indeed captured by the modified group delay function as in the case of the minimum phase group delay function shown in Figure 6.2 (d). A minimization of mean square error approach is used to find the optimal values for α, and γ. Let the minimum phase group delay function be denoted by $\tau_{min}(\omega)$ and the modified group delay function by $\tau_{mod}(\omega)$.The minimum phase group delay function $\tau_{min}(\omega)$ shown in Figure 6.2 (d) serves as a reference template and the modified group delay function $\tau_{mod}(\omega)$ is computed for

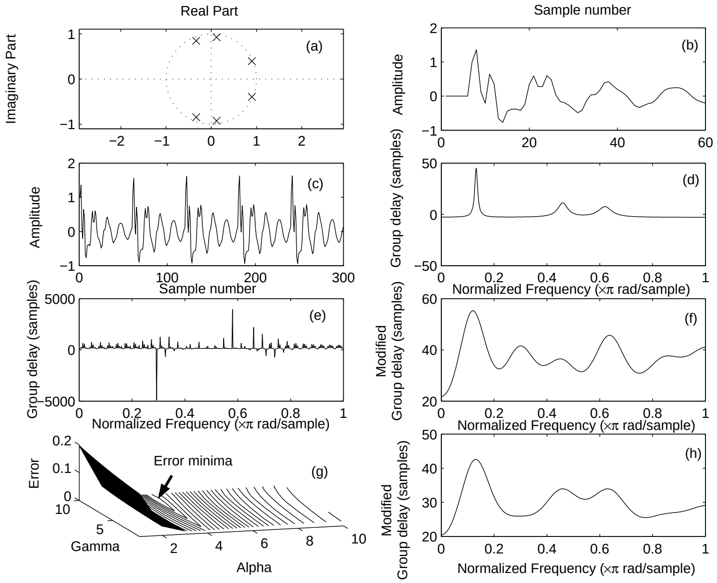

Fig. 6.2: Estimation of optimal $lifter_w$, α, and γ from a signal processing perspective. (a) z-plane plot of a system characterized by 3 formants (3 complex conjugate pole pairs), (b) Impulse response of the system shown in (a), (c) Response of the system in (a) excited with 5 impulses spaced 60 apart, (d) Group delay spectrum of the response in (a), (e) Group delay spectrum of the response in (c), (f) Modified group delay spectrum of the response in (c) for $lifter_w = 6$, $\alpha = 1$, and $\gamma = 1$, (g) Mean square error plot for α and γ (varied in steps of 0.1), (h) Modified group delay spectrum of the response in (d) for $lifter_w = 6$, $\alpha = 0.4$ and $\gamma = 0.9$.

various values of α, and γ. The parameters α and γ are varied in steps of 0.1 over the range 0 to 1. The mean square error (MSE) between $\tau_{mod}(\omega)$ and $\tau_{min}(\omega)$ is given by

$$MSE = \frac{1}{N}\sum_{1}^{N} e(k)^2 \tag{6.5}$$

where

$$e(k) = (\tau_{mod}(\omega) - \tau_{min}(\omega)) \tag{6.6}$$

and N is the length of $\tau_{mod}(\omega)$ or $\tau_{min}(\omega)$. The corresponding mean square error plot for α and γ (varied in steps of 0.1) over the range 0 to 1 is shown in Figure 6.2 (g). The error plot converges to a global minima at a value of $\alpha = 0.4$ and $\gamma = 0.9$. The error curve does not change for lengths of $lifter_w$ from 4 to 9. The envelope of the modified group delay spectrum for $\alpha = 0.4$ and $\gamma = 0.9$ is shown in 6.2 (h), and it is able to capture the formant information correctly.

6.3.3 Estimation of Optimal Values for $lifter_w$, α, and γ using Line Search

A series of experiments conducted initially showed that the values of the three parameters $lifter_w$, α, and γ have a large impact on the recognition error rate in all the three speech processing tasks mentioned earlier. Based on the results of these initial experiments we fix the length of the $lifter_w$ to 8 although the performance remains nearly the same for lengths from 4 to 9. Any value greater than 9 hurts performance badly. Having fixed the length of the $lifter_w$, the task now is to fix the values of α and γ. In order to estimate the values of α and γ an extensive optimization was carried out in [37] for the SPINE database [51] for phoneme recognition. To ensure that the optimized parameters were not specific to a particular database, we collected the sets of parameters that gave best performance on the SPINE database as in [37] and tested them on other databases like the DBIL database (for syllable recognition), TIMIT, NTIMIT (for speaker identification), and the OGI_MLTS database (for language identification). The values of the parameters that gave the best performance across all databases and across all tasks were finally chosen for the experiments. The optimization technique uses successive line searches. For each iteration, α is held constant, and γ is varied from 0 to 1 in increments of 0.1 (line search) and the recognition rate is noted for the three tasks on the aforementioned databases. The value of γ that maximizes the recognition rate is fixed as the optimal value. A similar line search is performed on α (varying it from 0 to 1 in increments of 0.1) keeping γ fixed. The set of values of α and γ that give the lowest error rate across the three tasks is retained. The series of experiments conducted to estimate the optimal values for $lifter_w$, α and γ using line search are summarized in Table 6.1. Based on the experiments conducted as in Table 6.1, the best front end across all tasks and across all databases used in this study is given in Table 6.2. In Tables 6.1 and 6.2, N_c is the number of cepstral co-efficients. It is emphasized here that the values of $lifter_w$, α, and γ listed in Table 6.2, are used for the evaluation of the MODGDF for all the three tasks namely speaker, language, and syllable recognition.

Table 6.1: Series of experiments conducted on various databases with the MODGDF.

Experiments conducted on the various databases
N_c = 10, 12, 13, 16
γ = {0.1 - 1.0} in increments of 0.1
α = {0.1 - 1.0} in increments of 0.1
$lifter_w$ = 4, 6, 9, 10, 12

Table 6.2: Best front-end for the MODGDF across all tasks and across all databases used in the study.

γ	α	$lifter_w$	N_c
0.9	**0.4**	**8**	**16**

6.4 BASELINE SYSTEM AND EXPERIMENTAL RESULTS FOR AUTOMATIC SPEAKER IDENTIFICATION

The baseline system used in this study uses the principle of likelihood maximization. A series of GMMs (See Appendix C) are used to model the voices of speakers for whom training data is available [34]. Single state, 64 mixture Gaussian mixture models (GMMs) are trained for each of the 600 speakers in the database. A classifier evaluates the likelihoods of the unknown speaker's voice data against these models. The model that gives the maximum accumulated likelihood is declared as the correct match. Out of the 10 sentences for each speaker, 6 were used for training, and 4 were used for testing. The tests were conducted on 600 speakers (600 x 4 tests) and the number of tests was 2400. A summary of results of performance evaluation for various features on both the TIMIT [3] (clean speech data) and NTIMIT [4] (noisy telephone data) corpora using GMM as the backend are listed in Table 6.3.

Table 6.3: Recognition performance of various features for speaker identification. MODGDF (MGD), MFCC (MFC), LFCC (LFC), spectral root compressed MFCC (SRMFC), energy root compressed MFCC (ERMFC), and spectral root compressed LFCC (SRLFC)

Task	**Feature**	**Database**	**Recognition**
Speaker Identification	MGD	TIMIT Clean speech	**99%**
	MFC		98%
	SRMFC		97.25%
	ERMFC		98%
	LFC		96%
	SRLFC		97%
Speaker Identification	MGD	NTIMIT Noisy telephone speech	**36%**
	MFC		34%
	SRMFC		34.25%
	ERMFC		34.75%
	LFC		30.25%
	SRLFC		31.75%

6.4.1 Discussion

For the TIMIT data the MODGDF gave a recognition performance of 99%. The performance of the MODGDF for this task is better than that of the spectral root compressed MFCC (root = 2/3) at 97.25%, the log compressed MFCC at 98% and energy root compressed MFCC (root = 0.08) at 98% as indicated in Table 6.3. For the NTIMIT data the MODGDF gave a recognition performance of 36%. The performance of the MODGDF for this task is better than that of the spectral root compressed MFCC (root = 2/3) at 34.25%, the log compressed MFCC at 34% and energy root compressed MFCC (root = 0.08) at 34.75% as indicated in Table 6.3. The performance of the the two forms of LFCCs are also listed in Table 6.3. It is emphasized again that the value of the root in both forms of the root compressed features has been taken after careful optimization using line search.

6.5 BASELINE SYSTEM AND EXPERIMENTAL RESULTS FOR LANGUAGE IDENTIFICATION

The baseline system used for this task is very similar to the system used for the automatic speaker identification task, except that each language is now modeled by a GMM (See Appendix C). Single state, 64 mixture Gaussian mixture models (GMMs) are trained for each of the 11 languages in the database. Out of 90 phrases for each language 45 were used for training and 20 were used for testing. The length of the test utterance was 45 seconds. The average recognition performance across 3 languages, for the 3 language task, and across 11 languages, for the 11 language task, is computed. A summary of the results of performance evaluation for various features on both DBIL and OGI_MLTS corpora using the GMM scheme are listed in Table 6.4.

6.5.1 Discussion

For the 3 language task on the DBIL database the MODGDF gave a recognition performance of 96%. The performance of the MODGDF for this task is better than that of the spectral root compressed MFCC (root = 2/3) at 95%, the log compressed MFCC at 95% and energy root compressed MFCC (root = 0.08) at 95.4% as indicated in Table 6.4. For the 11 language task on the *OGI_MLTS* data the MODGDF gave a recognition performance of 53%. The performance of the MODGDF for this task is better than that of the spectral root compressed MFCC (root = 2/3) at 50.4%, the log compressed MFCC at 50% and energy root compressed MFCC (root = 0.08) at 50.6% as indicated in Table 6.4. The performance of the the two forms of LFCCs are also listed in Table 6.4. Once again it is emphasized that the value of the root in both forms of the root compressed features has been taken after optimization using line search.

Table 6.4: Recognition performance of various features for language identification. MODGDF (MGD), MFCC (MFC), LFCC (LFC), spectral root compressed MFCC (SRMFC), energy root compressed MFCC (ERMFC), and spectral root compressed LFCC (SRLFC)

Task	**Feature**	**Database**	**Recognition**
Language Identification	MGD	DBIL Broadcast News 3 Language Task	**96%**
	MFC		95%
	SRMFC		95%
	ERMFC		95.4%
	LFC		92%
	SRLFC		92.6%
Language Identification	MGD	OGI_MLTS Telephone speech 11 Language Task	**53%**
	MFC		50%
	SRMFC		50.4%
	ERMFC		50.6%
	LFC		47%
	SRLFC		48%

6.5.2 Confusability Analysis for the 11 Language Task

The confusion matrix for the 11 language task using the MODGDF is shown in Table 6.5. The matrix is constructed from the results of testing 20 utterances (18 male and 2 female) each of duration 45 seconds from the OGI_MLTS database. From the matrix it significant to note that English and French are easily discriminated from each other and also from the other languages. Confusability between Japanese and Korean is very high. In order to cross check the confusability between languages, vector quantization (VQ) codebooks of size 32, representing the languages English, French, Korean, and Japanese are computed. For each language, we now have a VQ codebook of size 32, where each codevector is 13-dimensional. The 13-dimensional VQ codebooks are reduced to 2-dimensional codebooks using the Sammon mapping technique as already described in Chapter 5. The visualizations of the 2-dimensional codebooks of English-French are shown in Figure 6.3, and Korean-Japanese are shown in Figure 6.4, respectively. It is evident from these illustrations that the languages English and French are well separated by the MODGDF while there is a high degree of overlap between Korean and Japanese. It is worthwhile to note from the confusion matrix in Table 6.5, that the language pair English-French are least confused while

Table 6.5: Confusion matrix for the 11 language task, E: English, Fr: French, K: Korean, M: Mandarin, Fa: Farsi, G: German, S: Spanish, H: Hindi, V: Vietnamese, T: Tamil and J: Japanese.

	E	*Fr*	*K*	*M*	*Fa*	*G*	*S*	*H*	*V*	*T*	*J*
E	16	0	3	0	0	1	0	0	0	0	0
Fr	0	16	2	0	0	0	0	2	0	0	0
K	2	0	10	0	0	0	0	0	1	0	7
M	3	0	4	7	0	3	0	0	0	0	3
Fa	5	0	0	0	11	3	0	0	1	0	0
G	7	0	2	0	3	8	0	0	0	0	0
S	2	2	0	0	0	0	13	0	3	0	0
H	5	0	0	0	2	1	0	6	3	3	0
V	3	0	1	1	1	2	0	0	8	4	0
T	3	0	0	0	3	0	0	2	1	11	0
J	0	2	9	0	2	0	0	0	2	0	5

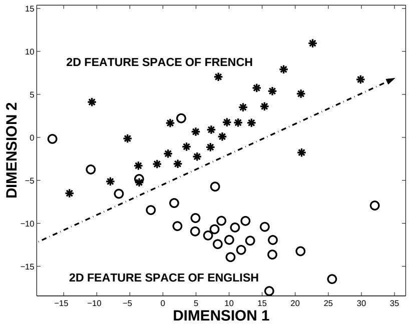

Fig. 6.3: English-French language discrimination with the MODGDF.

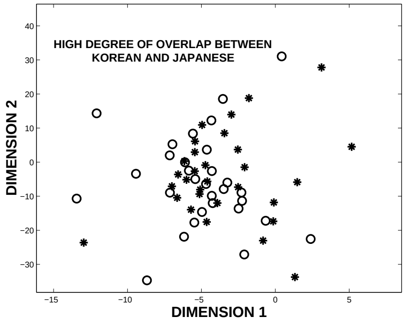

Fig. 6.4: Korean-Japanese language discrimination with the MODGDF.

the language pair Korean-Japanese are highly confused.

6.6 BASELINE SYSTEM AND EXPERIMENTAL RESULTS FOR SYLLABLE RECOGNITION

Conventional speech recognition systems use a large vocabulary lexicon and highly perplex language models to achieve good recognition rates. The primary reason for adopting such an approach is because the baseline recognition techniques fail to align phonetic unit boundaries accurately. A novel method of segmentation of continuous speech into syllable like units, using minimum phase group delay functions derived from the root compressed short-time energy function has been discussed in [14], and [52]. These speech segments are then recognized in isolated style, using hidden Markov models (See Appendix D), and concatenated in the same sequence as they were segmented. Once the recognition of a particular phrase of speech is completed, a forced tri-syllabic Viterbi re-alignment technique for improving the recognition performance of a continuous speech recognizer developed using such an approach. The proposed technique uses the logic of syllable string optimization using a forced tri-syllabic Viterbi re-alignment. This technique addresses the issue of reduced recognition performance due to errors at the segmentation stage in a segment-based recognition system. The MODGDF is used as a front end for the recognition of two Indian languages namely Tamil and Telugu on the DBIL database [5], using the aforementioned baseline system. The recognition results are presented and its implications discussed at the end of this Section.

6.6.1 Overview of The Segment-Based Approach to Recognition of Continuous Speech

In a segment-based approach to recognition of continuous speech, the test phrase is segmented at boundaries of syllabic units using the minimum phase group delay function derived from the root compressed short-time energy function [14, 52]. These segments are then checked in isolated style against all HMMs built apriori, and concatenated in

the same sequence as they were segmented. Figure 6.5, demonstrates the segmentation of one phrase from the Telugu Doordarshan news bulletin [5], using manual, short-time energy based, and group delay based methods. Figure 6.5 (a), denotes a signal for the speech utterance */mukhya mantri chandrabAbu nAyuDu AdEshham/* and Figure 6.5 (b), is the corresponding short term energy function. Figure 6.5 (c), shows the minimum phase group delay function. The thick vertical lines in Figure 6.5 (c) which pass

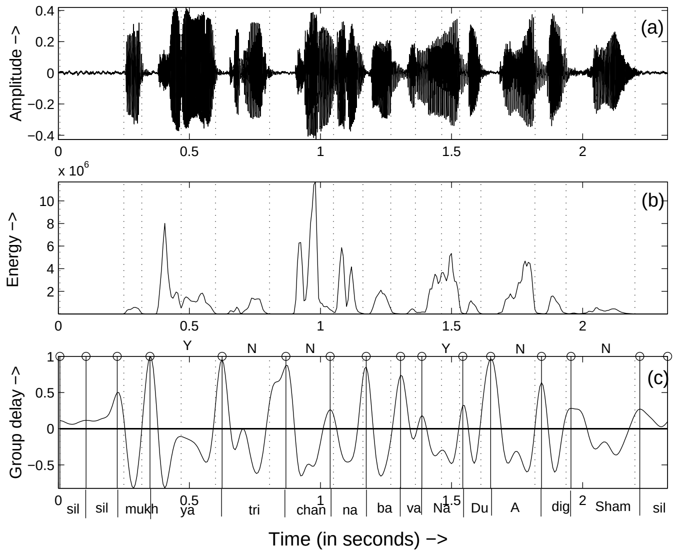

Fig. 6.5: Comparison of segmentation of continuous speech using two methods. (a) Speech signal for Telugu speech utterance */mukhya mantri chandrabAbu nAyuDu AdEshham/*, (b) its short-time energy function, (c) minimum phase group delay function derived from short-time energy function. Thick vertical lines and dotted vertical lines denote the identified segment boundaries and the actual segment boundaries, respectively. Syllables at the bottom of (c) denotes the recognized syllables.

through the peaks of the minimum phase group delay function denote the segment boundaries obtained by the automatic segmentation algorithm [14], the dotted vertical lines correspond to manually found segment boundaries of the speech signal and the transcription below denotes the recognized syllable for the corresponding segment.

6.6.2 Issues in Segment-Based Recognition of Continuous Speech

A test utterance from the Telugu corpus which reads */I rOju vArtallOni mukhyAmshAlu/* is considered, to illustrate two major issues in a segment based continuous speech recognition system. The utterance is segmented at syllabic boundaries and each segment is tested in isolated style against HMMs built apriori as mentioned in Section 6.6.1. The baseline system outputs the first best alternative based on the maximum likelihood value from among a set of *n*-best alternatives. Let us consider the five best alternatives of each of the segments from the above mentioned utterance as in Table 6.6. In Table 6.6, the first column contains the actual syllable, while each row contains the first five alternatives in the decreasing order of their likelihood values. The highlighted syllable in each row denotes the correct match with the actual syllable. In this context, the two major issues in a segment based continuous speech recognition system are

- In many instances the actual syllable is one among the first few alternatives or positions, if it is not in the first position. For example the actual syllable */mukh/* in the 9^{th} row in Table 6.6, is not in the first position but in the 4^{th} position.

- If the speech signal is segmented at correct syllabic boundaries and the recognition is wrong, then the recognized syllable is phonetically very near to the actual syllable. For example in row 6, of Table 6.6, */tal/* has been identified as */tam/*.

Table 6.6: Five best alternatives of each recognized segment for the utterance */I rOju vArtallOni mukhyAmshAlu/* using the baseline system.

Actual Syllable	1	2	3	4	5
SIL	**SIL**	i	chep	te	a
I	**I**	i	E	yE	lI
rO	**rO**	lu	Du	ru	gu
ju	**ju**	ja	du	Du	rU
vAr	**vAr**	vA	var	va	van
tal	tam	Ta	tun	ta	Tar
lO	**lO**	lA	dra	ga	dA
ni	**ni**	nI	nE	I	vI
mukh	mu	mup	ku	**mukh**	mU
yAm	ya	Tam	yam	Ta	yar
shA	**shA**	sha	sA	sai	sham
lu	li	**lu**	I	yu	yO

6.6.3 Tri-Syllabic Viterbi Re-Alignment

To address the two major issues in a segment based continuous speech recognition system an approach based on tri-syllabic Viterbi re-alignment is used. The technique is described by the following steps.

- The HMM decoder is used to generate *5*-best likely alternative syllabic units corresponding to each isolated test segment.
- All possible permutations with these *5* basic units at each position is enumerated and a list of possible syllable strings (sequences) is found. The correct syllable string (sequence) should be one among the list of possible syllable strings (sequences).

- The problem of finding the string of syllabic units with maximum likelihood is now transformed to the problem of finding an optimal state sequence.

The test phrase is first recognized using the segment based continuous speech recognition system. To implement the *tri-syllabic Viterbi re-alignment*, three consecutive syllables from the recognized phrase, with five possible alternatives to each syllable, are considered. This can now be viewed as a concatenation of HMMs in parallel, where each HMM corresponds to a syllable string of three syllables, which is called the *t*ri-syllable. This is equivalent to applying Viterbi re-alignment locally over a set of three syllables (tri-syllable). The re-alignment is done in sets of three syllables across the length of the phrase and is therefore called the tri-syllabic Viterbi re-alignment. After the tri-syllabic Viterbi re-alignment and the succeeding recognition process the middle syllable is picked as the final recognized syllable. This technique is applied across the entire phrase. To illustrate the tri-syllabic Viterbi re-alignment, we consider A typical network for a particular tri-syllabic string/ *ni mukh yAm* / [1] of Telugu language. The parallel network consisting of various possible tri-syllabic strings as a parallel concatenation of HMMs, is shown in Figure 6.6. Note that a check mark is placed against the right syllable sequence in the network. The tri-syllabic Viterbi re-alignment technique based on the hypothesis that the syllable */mukh/* which lies in the fourth place in a 5-best result, as listed in Table 6.6, pushes it to the first place when a forced re-alignment is performed. The recognition result of one phrase from the Tamil news corpus where the baseline system wrongly identifies two syllables and the tri-syllabic Viterbi re-alignment technique is able to correct them is illustrated below.

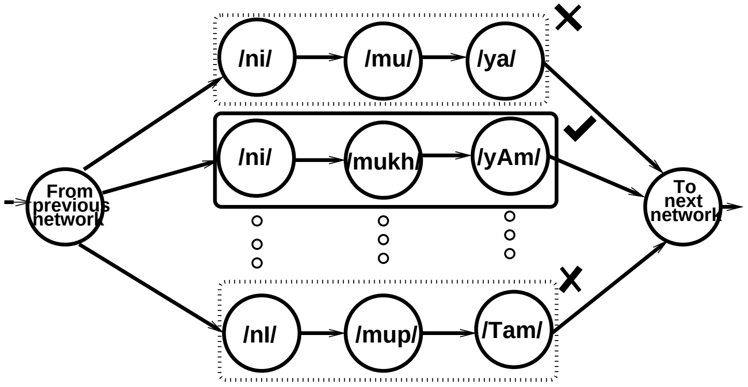

Fig. 6.6: Parallel network for a particular syllable string/*ni mukh yAm*/ of Telugu language.

---

[1] This tri-syllabic string is a part of the phrase */I rOju vArtallOni mukhyAmshAlu/* illustrated in Table 6.6

Original syllable sequence:

muQn QnAL mu da la maic car sel vi je ya la li tA

Recognized syllable sequence of the baseline system:

muQn QnAL mu var lA mai ci su Lai vi je ya rar du tA SIL

Recognized syllable sequence after local forced viterbi realignment:

muQn QnAL mu da lA maic ci sel Lai vi je ya la du tA SIL

The syllables /da/, /maic/, /sel/ and /la/ which are wrongly identified by the baseline system are corrected by the enhanced system with local forced viterbi realignmnent. It is indeed true that the Viterbi re-alignment process could introduce errors locally in some cases, when the automatic segmentation has succeeded in marking the exact syllable boundaries. But the technique itself is based on the fact that on an average the re-alignment process improves the overall recognition performance. It is significant to note that, this technique does not use highly perplex language models to re-score sentences according to grammar or semantics, to address errors in segmentation. But it does re-score the recognized phrase coming from the segment-based recognition system by applying tri-syllabic Viterbi re-alignment locally. This makes the system more implicit, less dependent on linguistic information and poses enormous new possibilities.

6.6.4 Extension of The Tri-Syllabic Viterbi Re-alignment to 15-best Alternatives

The application of the tri-syllabic Viterbi re-alignment on 5-best alternatives has been discussed in Section 6.6.2. In an effort to further improve the recognition performance, the tri-syllabic Viterbi re-alignment is applied on the 15-best alternatives computed from the baseline system. The hypothesis in considering the 15-best alternatives is that the correct syllable may sometimes be lying in a position greater than the 5^{th} or the 10^{th} position and the tri-syllabic Viterbi re-alignment should be able to push it to the first position. To illustrate the significance of applying the tri-syllabic Viterbi re-alignment on 15-best alternatives, we consider one phrase

/ SIL dE si ya vA da kAZN gras /

from the Tamil news database. The baseline recognition recognizes it as

/ SIL dE si Nu la luk kAZN muQn gras /

Note that the syllables recognized in-correctly are /Nu/, /la/, and /luk/. The syllable **/muQn/** is an insertion error. The phrase is recognized on applying the tri-syllabic Viterbi re-alignment on 5-best alternatives as

/ SIL dE si ya la da kAZN ni gras /

It is significant to note that the re-alignment has succeeded in correcting two errors corresponding to the syllables /ya/ and /da/. But it fails to correct the error in the syllable which is recognized as /la/ instead of /vA/. It is significant to note that the correct syllable /vA/ is not present in the 5-best alternatives for the particular phrase under question as illustrated in Table 6.7[2]. In Table 6.7, the first column contains the actual syllable, while each row contains the first five alternatives in the decreasing order of their likelihood values. The highlighted syllable in each row denotes the correct match withthe actual syllable. It is clear that the tri-syllabic Viterbi re-alignment fails to correct the syllable /vA/ which is recognized in-correctly primarily because the 5-best alternatives do not contain the syllable at all. When the tri-syllabic Viterbi re-alignment is extended to 15-best alternatives the phrase under question is recognized as **/ SIL dE si ya vA da kAZN ni gras /**

---

[2] Note that the insertion error is not shown in Table 6.7.

Table 6.7: Five best alternatives of each recognized segment for the utterance */SIL dE si ya vA da kAZN gras/* using the baseline system.

Actual Syllable	1	2	3	4	5
SIL	**SIL**	su	dum	Ki	Ra
dE	**dE**	je	vE	Sey	Cey
si	**si**	ru	Li	Vi	Ri
ya	Nu	vu	lu	Qna	**ya**
vA	la	lA	a	Du	tak
da	Luk	du	**da**	Du	ru
kAZN	**kAZN**	A	dA	Du	ru
gras	**gras**	ru	da	yum	mi

Note that the syllable vA which is recognized in-correctly by applying tri-syllabic Viterbi re-alignment on 5-best alternatives is corrected when a 15-best alternatives are considered. In this context it is also significant to note that an extension of the forced tri-syllabic Viterbi re-alignment technique to n-best alternatives can further improve recognition performance.

6.6.5 The Baseline System

The baseline recognition system uses Hidden Markov Models trained apriori for 320 syllables for Tamil and 265 syllables for Telugu. The number of syllables used for training are selected based on their frequency of occurrence in the respective corpora. During the training phase, Hidden Markov Models (HMMs) are built for every syllable that occurs more than 50 times in the corpus. A separate model is built for silence. 5 state HMM's with 3 mixtures/state are used throughout the experimental study. The reduced vocabulary size as a fraction of the original vocabulary size, as used in the baseline system is depicted as a pie chart in Figure 6.7. During the testing phase

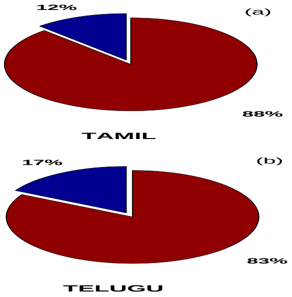

Fig. 6.7: Reduced Vocabulary size Tamil and Telugu. (a) Tamil and (b) Telugu.

the test sentence is segmented at boundaries of syllabic units using the minimum phase group delay function derived from the causal portion of the root compressed energy function assuming that it is an arbitrary magnitude spectrum exactly as in [14]. These segments are now checked in isolated style against all HMMs built apriori. The HMM that gives the maximum likelihood value is declared as the correct match. The recognized isolated syllables are now concatenated in the same order as they appear in the test sentence to output the recognized sentence. The forced tri-syllabic Viterbi re-alignment technique as discussed in Section 6.6.3 is used on the recognized phrase. The recognition results using the baseline system with a forced tri-syllabic Viterbi re-alignment technique on 15-best alternatives for two news bulletins each of duration fifteen minutes, comprising 9400 syllables for Tamil and Telugu are illustrated in Table 6.8.

6.6.6 Discussion

For the Telugu data the MODGDF gave a recognition performance of 38.2% for whole syllable recognition. The performance of the MODGDF for this task is better than that of the spectral root compressed MFCC (root = 2/3) at 35.6%, energy root com-

Table 6.8: Recognition performance of various features for syllable recognition. MODGDF (MGD), MFCC (MFC), LFCC (LFC), spectral root compressed MFCC (SRMFC), energy root compressed MFCC (ERMFC), and spectral root compressed LFCC (SRLFC)

Task	Feature	Database	Recognition
Syllable Recognition	MGD	DBIL Telugu broadcast news	**38.2%**
	MFC		38.6%
	SRMFC		35.6%
	ERMFC		38%
	LFC		32.6%
	SRLFC		34.2%
Syllable Recognition	MGD	DBIL Tamil broadcast news	**36.7%**
	MFC		37.1%
	SRMFC		34.1%
	ERMFC		36.5%
	LFC		31.2%
	SRLFC		32.4%

pressed MFCC (root = 0.08) at 38%, slightly less than the log compressed MFCC at 38.6% as indicated in Table 6.8. For the Tamil data the MODGDF gave a recognition performance of 36.7% for whole syllable recognition. The performance of the MODGDF for this task is better than that of the spectral root compressed MFCC (root = 2/3) at 34.1%, energy root compressed MFCC (root = 0.08) at 36.5% slightly less than the log compressed MFCC at 37.1% as indicated in Table 6.8. The performance of the the two forms of LFCCs are also listed in Table 6.8. It is emphasized here that the value of the root in both forms of the root compressed features has been selected after careful optimization using line search. In the following Section, the issues involved in combining the MODGDF with various other features, is discussed extensively.

6.7 SIGNIFICANCE OF FEATURE COMBINATIONS

The technique of combination is widely used in statistics. The simplest method of combination involves averaging the various estimates of the underlying information. This idea is based on the hypothesis that if different estimates are subject to different sources of noise then combining them will cancel some of the errors when an averaging is done. A good example of combining features is the case of Janin [53], who have trained neural networks based on different random starting conditions to combine different features after the acoustic model. A combination system works on the principle that if some characteristics of the speech signal that are de-emphasized by a particular feature are emphasized by another feature, then the combined feature stream captures complementary information present in individual features.

6.7.1 Feature Combination Before the Acoustic Model

The combination of features before the acoustic model have been used by Okawa [54] and Ellis [55], where efforts have been made to capitalize on the differences between various feature streams using all of them at once. The joint feature stream is derived in such an approach by concatenating all the individual feature streams into a single feature stream.

6.7.2 Likelihood Combination After the Acoustic Model

This approach uses the technique of combining the outputs of the acoustic models. Complex techniques of combining the posteriors [53,56–58] have evolved but the classic way of simple averaging of the maximum likelihoods from different estimators is the best approach to feature combination after the acoustic model. In this context, it is also worthwhile to note that if the intent is to capitalize on the complementary information in different features the posteriors of the same classifier for individual features can be combined to achieve improved speech recognition performance.

6.8 RECONSTRUCTION OF FORMANTS FROM THE MODGDF, MFC AND JOINT FEATURES

In this Section, we reconstruct the formant structures or the respective short time spectra from the MODGDF, MFCC, and joint features. The MODGDF is derived from the modified group delay spectra as

$$c_p(n) = \sum_{k=0}^{k=N_f} \tau_m(k) \cos(n(2k+1)\pi/N_f) \qquad (6.7)$$

where N_f is the DFT order. It is emphasized here that there are no filter banks used in the computation of the MODGDF. The MFCC are derived from the short time power spectra as

$$c_m(n) = \sum_{k=1}^{k=N_{fb}} X_k \cos(n(k-\frac{1}{2})\pi/N_{fb}) \qquad (6.8)$$

where $n = 1, 2, 3, ...M$, represents the number of cepstral coefficients and $k = 1, 2, 3, ...$ Note that N_{fb} the number of filter banks used. X_k represents the log energy output of the k^{th} filter. The joint features (MODGDF + MFCC) are derived by appending the MODGDF vectors calculated as in Equation 6.7 with the MFCC vectors calculated as in Equation 6.8. The number of cepstral co-efficients used in both the MODGDF and the MFCC are the same. To reconstruct the formant structures or the short-time spectra from the cepstra an inverse DCT of the original DFT order has to be performed on the cepstra. The reconstructed modified group delay spectra as derived from the MODGDF is given by

$$\tau_m(k) = \sum_{n=0}^{n=N_f} c_p(n) \cos(n(2k+1)\pi/N_f) \qquad (6.9)$$

where N_f is the DFT order. The reconstructed short time power spectra derived from the MFCC is given by

$$X_k = \sum_{n=0}^{n=N_f} c_m(n) \cos(n(2k+1)\pi/N_{fb}) \qquad (6.10)$$

where $n = 1, 2, 3, ...M$, represents the original number of cepstral coefficients and $k = 1, 2, 3, ...N_{fb}$, where N_{fb} is the original number of filter banks used. X_k represents the reconstructed log energy output of the k^{th} filter. The smooth frequency response of the original DFT order is computed by interpolating the filter bank reconstructed energies.

6.8.1 Algorithm for Reconstruction of the Short-Time Power Spectra from the MFCC

The algorithm for the reconstruction of the short time power spectra from the MFCC is enumerated below

1. Start with the n dimensional cepstral feature vector (MFCC).
2. Select the number of filters in the filter bank equal to N_{fb}, the same number as used in the forward discrete cosine transformation.
3. Generate a matrix D', which is a $\{N_{fb} \times n\}$ DCT matrix. This DCT matrix is the transpose of the matrix D, which is a $\{n \times N_{fb}\}$ DCT matrix, used in the forward transformation. The forward DCT matrix is generated using the definition of the forward DCT as in Equation 6.8.
4. Compute the filter bank reconstructed spectra by taking a product of the matrix D', which is a $\{N_{fb} \times n\}$ DCT matrix, and the matrix C, which is a $\{n \times f_n\}$ cepstral matrix, where f_n is the number of frames into which the speech signal is frame blocked. The filter bank reconstructed spectral matrix F_b thus derived is a $\{N_{fb} \times f_n\}$ matrix, where N_{fb} is the number of filter banks and f_n is the number of frames into which the speech signal is frame blocked.
5. Invert the filter bank center frequencies and compute both the integer and fractional sampling positions by interpolation. These sampling positions give the exact position of each FFT bin in the filter bank.
6. Reconstruct the smoothed short-time power spectra of the original DFT order by using the sampling positions as computed in step 5.

6.8.2 Algorithm for Reconstruction of Short-Time Modified Group Delay Spectra from the MODGDF

The algorithm for the reconstruction of the short time modified group delay spectra from the MODGDF is enumerated below

1. Start with the n dimensional cepstral feature vector (MODGDF).
2. Select the DFT order as N_f, the same order as used in the forward transformation (DFT order is 512 here). Note that there are no filter banks involved in the computation of the MODGDF.
3. Generate a matrix D', which is a $\{N_f \times n\}$ DCT matrix. This DCT matrix is the transpose of the matrix D, which is a $\{n \times N_f\}$ DCT matrix, used in the forward transformation. The forward DCT matrix is generated using the definition of the forward DCT as in Equation 6.7.
4. Compute the reconstructed short-time modified group delay spectraby taking a product of the matrix D', which is a $\{N_f \times n\}$ DCT matrix, and the matrix C, which is a $\{n \times f_n\}$ cepstral matrix, where f_n is the number of frames into which the speech signal is frame blocked. The reconstructed short time modified group delay spectral matrix MGD_{recon} thus derived is a $\{N_f \times f_n\}$ matrix, where N_f is the original DFT order and f_n is the number of frames into which the speech signal is frame blocked.

6.8.3 Algorithm for Reconstruction of Short-Time Composite Spectra from Joint Features (MODGDF+MFCC)

1. Reconstruct the short time power spectra of the original DFT order from the n dimensional MFCC as in Section 6.8.1
2. Reconstruct the short time modified group delay spectra of the original DFT order from the n dimensional MODGDF as in Section 6.8.2
3. Add the short-time power spectra reconstructed from the MFCC and the short-time modified group delay spectra reconstructed from the MODGDF to compute

the short time composite spectra of the original DFT order. Note that the dimensionality of the MODGDF and the MFCC is the same.

6.8.4 Formant Reconstruction for a Synthetic Vowel

Typically a vowel is characterized by the first three formants. Assuming a source system model of speech production, the transfer function of such a system is given by

$$H(z) = \frac{1}{1 - 2e^{-\pi B_i T} cos(\omega_i T) z^{-1} + e^{-2\pi B_i T} z^{-2}} \tag{6.11}$$

In the above array of equations ω_i is the radian frequency of the i^{th} formant, B_i is the bandwidth of the i^{th} formant, and T is the sampling period. Using equation 6.11 we generate a synthetic vowel with the following values F1 = 500Hz, F2 = 1500Hz, F3 = 3500Hz, B_i = 10% of F_i, and T = 0.0001s corresponding to a sampling rate of 10 KHz. We then extract the MODGDF, MFCC and joint features (MODGDF + MFCC) from it. To reconstruct the formants we use the algorithms discussed in earlier Sections. The reconstructed formant structures derived from the MODGDF, MFCC,joint features (MODGDF + MFCC) and also RASTA filtered MFCC are shown in Figures 6.8 (a), (b), (c), and (d) respectively. The illustrations are shown as spectrogram like plots where the data along the y-axis correspond to the DFT bins and the x-axis corresponds to the frame number. The y-axis in DFT bins can be converted to Hertz as

$$Frequency_{DFTBINS} = F_s/N \tag{6.12}$$

where F_s is the sampling rate and N is the number of points over which the DFT is computed. In Figure 6.8, the sampling rate F_s = 10 KHz and N = 512. It is interesting to note that in Figure 6.8 (a), all the 3 formants are visible. In Figure 6.8 (b), while the first 2 formants visible, the 3rd formant is not clearly visible. In Figure 6.8 (c) while first 2 formants are emphasized further, the 3rd formant is also visible. Hence it is clear that joint features are able to combine the information that is available in both the features. The RASTA filtered MFCC shown in Figure 6.8 (d) fails to capture the formant structure for the synthetic vowel. Similar spectrogram like

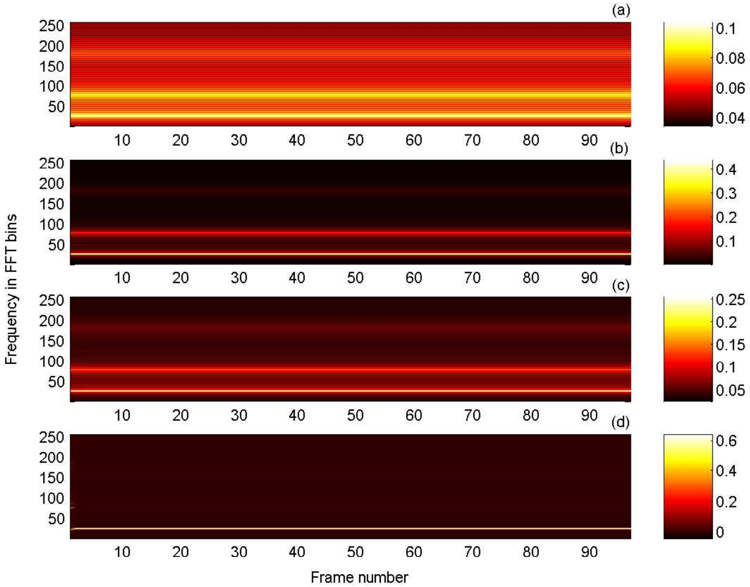

Fig. 6.8: Spectrogram like plots to illustrate formant reconstructions for a synthetic vowel. (a) The short-time modified group delay spectra reconstructed from MODGDF (b) The short-time power spectra reconstructed from MFCC, (c) The short-time composite spectra reconstructed from joint features (MODGDF+MFCC), and (d) The short-time power spectra reconstructed from RASTA filtered MFCC.

plots to illustrate formant reconstructions for a synthetic speech signal with varying formant trajectories is shown in Figure 6.9. It is interesting to note that in Figure 6.9 (a), all the 3 formants are visible. In Figure 6.9 (b), while the first 2 formants visible, the 3rd formant is not clearly visible. In Figure 6.9 (c) while first 2 formants are clear, the 3rd formant is further emphasized. Hence it is clear that joint features are able to combine the information that is available in both the features.

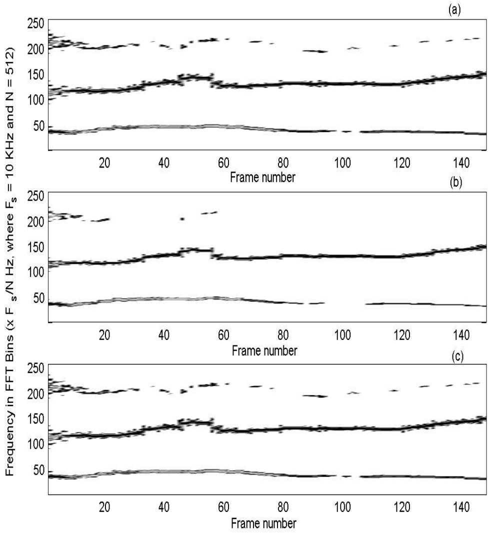

Fig. 6.9: Spectrogram like plots to illustrate formant reconstructions for a synthetic speech signal with varying formant trajectory. a) The short-time modified group delay spectra reconstructed from the MODGDF b) The short-time power spectra reconstructed from MFCC, and c) The short-time composite spectra reconstructed from joint features (MODGDF+MFCC)

6.8.5 Formant Reconstruction for Clean and Noisy Speech Data

In this Section, we extract the MODGDF, MFCC and joint features (MODGDF + MFCC) for both clean speech and speech corrupted by car noise at a SNR of 2 dB. The reconstructed formant structures derived from the MODGDF, MFCC, and joint features (MODGDF + MFCC) from clean speech are shown in Figures 6.10 (a), (c), and (e) respectively. The reconstructed formant structures derived from the MODGDF, MFCC, and joint features (MODGDF + MFCC) for the same segment of clean speech corrupted with car noise at a SNR of 2 dB are shown in Figures 6.10 (b), (d), and (f)

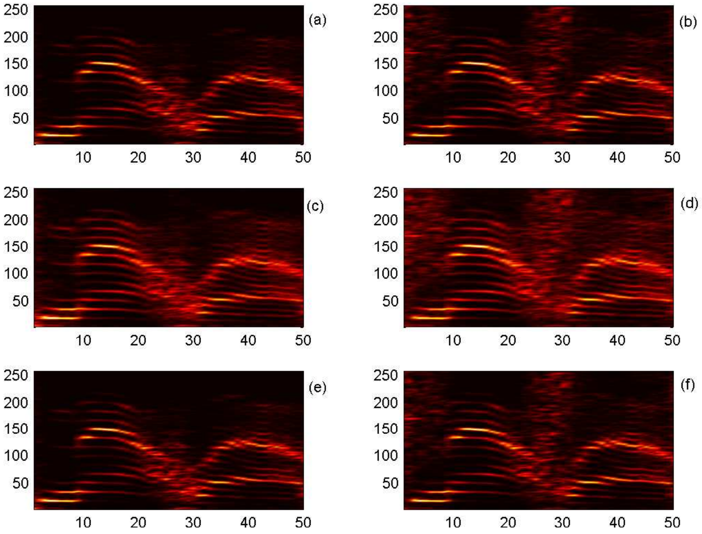

Fig. 6.10: Spectrogram like plots to illustrate formant reconstructions for clean and noisy speech. (a) The reconstructed short-time modified group delay spectra for clean speech, (b) The reconstructed short-time modified group delay spectra for the same segment of speech corrupted with car noise at a SNR of 2 dB, (c) The reconstructed short time power spectra (MFCC) for clean speech, (d) The reconstructed short time power spectra (MFCC) for the same segment of speech corrupted with car noise at a SNR of 2 dB, (e) The reconstructed short time composite spectra from joint features (MODGDF+MFCC) for clean speech, (f) The reconstructed short time composite spectra from joint features (MODGDF+MFCC) for the same segment of speech corrupted with car noise at a SNR of 2 dB.

respectively. The segment of speech corresponds to the word *"matlab"* uttered by a female speaker. The illustrations are shown as spectrogram like plots where the data along the y-axis correspond to the DFT bins and the x-axis corresponds to the frame number. From the spectrogram like plots of the reconstructed short time spectra, it is evident that the MODGDF is able to eliminate some spectral distortions due to noise as in Figure 6.10 (b), when compared to the MFCC as in Figure 6.10 (d). It is also interesting to note that the reconstructed short time spectra derived from the joint features for both clean and noisy speech shown in 6.10 (e) and (f) respectively, combine the information gathered by the individual features namely, the MODGDF and MFCC.

6.9 EXPERIMENTAL RESULTS FOR COMBINED FEATURES

In this Section, the results of performance evaluation of the MODGDF, MFCC, LFCC, and joint features, for four speech prcocessing tasks namely phoneme, syllable, speaker, and language recognition. We have also explored combining model based features like the LPCC with the MODGDF. But we present the results of combining MFCC with the MODGDF as this combination gave the best results among all the other combinations. The results of combining any two features derived from the short-time power spectra like the MFCC and the LFCC are also listed.

6.9.1 Extraction of Joint Features Before the Acoustic Model

The following method is used to to derive joint features by combining features before the acoustic model

- Compute 13 dimensional MODGDF and the MFCC streams appended with velocity, acceleration and energy parameters.
- Use Feature stream combination to append the 42 dimensional MODGDF stream to the 42 dimensional MFCC stream to derive a 84 dimensional joint feature stream.

Henceforth we use the subscript $_{bm}$ for joint features thus derived.

6.9.2 Likelihood Combination After the Acoustic Model

The following method is used to perform likelihood combination after the acoustic model

- Compute 13 dimensional MODGDF and the MFCC streams appended with velocity, acceleration and energy parameters.
- Build a Gaussian mixture model (GMM) (for phoneme, speaker, and language recognition tasks) or a Hidden Markov model (HMM) (for the continuous speech recognition task).
- Combine the posterior probability outputs derived using each model by weighting the probabilities appropriately.
- Make a decision based on maximization of the combined output probability.

Henceforth we use the subscript $_{am}$ for joint features thus derived.

6.9.3 Significance of Combining Other Features

To enable a fair comparison between combining the MODGDF and MFCC and combining other features, conventionally used in speech recognition, the following combinations have been tried

- The MODGDF and LFCC.
- The MFCC and LFCC.
- The MODGDF and LPCC.
- The MFCC and LPCC.

It was found from the experiments conducted that the best combination was that of the MODGDF and the MFCC in terms of recognition performance across all tasks. This is primarily because the MODGDF are derived from the short-time phase spectra and MFCC and LFCC are derived from the short-time power spectra. The LPCC as a feature performed poorly when compared to all the individual features and combining the LPCC with any other feature like the MFCC and LFCC gave no improvement in recognition performance. It is emphasized here that among the two combinations namely MFCC+LPCC and MFCC+LFCC, the combination of MFCC+LFCC gave the best improvement in recognition performance. Hence we list the recognition performance of combining the MODGDF and MFCC and also the recognition performance of combining the MFCC and LFCC.

6.9.4 Experimental Results of Combined Features for Speaker Identification

In this Section, the experimental results for speaker identification on the TIMIT database (clean speech) and the NTIMIT database (noisy telephone speech) using combined features are described. The experimental setup is similar to that described in Section 6.4 for speaker identification. Table 6.9, lists the results for speaker identification using individual and combined features. For the TIMIT (clean speech) data [3] the MODGDF (MGD) recognition performance was at 99%, MFCC (MFC) at 98%, ($\{\mathbf{MGD}+\mathbf{MFC}\}_{\mathbf{bm}}$) at 99.5%, and ($\{\mathbf{MGD}+\mathbf{MFC}\}_{\mathbf{am}}$) at 99.5% for this task. The best increase due to feature combination was 0.5%. While for the NTIMIT (noisy telephone speech) data [4] the MODGDF (MGD) recognition performance was at 36%, MFCC (MFC) at 34%, ($\{\mathbf{MGD}+\mathbf{MFC}\}_{\mathbf{bm}}$) at 42%, and ($\{\mathbf{MGD}+\mathbf{MFC}\}_{\mathbf{am}}$) at 40% for this task. The best increase due to feature combination was 6% as indicated in Table 6.9. To enable a fair comparison the MFCC was also combined with LPCC and LFCC. While combining the MODGDF with the LPCC, brought down recognition performance, the combined pair of MODGDF and LFCC gave a small improvement in recognition performance. The results for combining the MODGDF with the LFCC are also tabulated in Table 6.9.

Table 6.9: Recognition performance of combined features for speaker identification.

Task	**Feature**	**Database**	**Recogn.**	**Increase in Recogn.**
Speaker Identification	MGD	TIMIT Clean speech	**99%**	-
	MFC		98%	-
	LFC		96%	-
	$\{MGD+MFC\}_{bm}$		99.5%	**0.5%**
	$\{MGD+MFC\}_{am}$		99.5%	**0.5%**
	$\{MFC+LFC\}_{bm}$		98%	**0%**
	$\{MFC+LFC\}_{am}$		98%	**0%**
Speaker Identification	MGD	NTIMIT Telephone speech	**36%**	-
	MFC		34%	-
	LFC		30.25%	-
	$\{MGD+MFC\}_{bm}$		42%	**6%**
	$\{MGD+MFC\}_{am}$		40%	**4%**
	$\{MFC+LFC\}_{bm}$		34%	**0%**
	$\{MFC+LFC\}_{am}$		34%	**0%**

6.9.5 Experimental Results of Combined Features for Language Identification

In this Section, the experimental results for language identification on the DBIL database (3 language task) and the OGI_MLTS database (11 language task for noisy telephone speech) are listed. The experimental setup is similar to that described in Section 6.5, for language identification. The results of the MODGDF and the MFCC on both the DBIL and OGI_MLTS [47] corpora using the GMM scheme are listed in Table 6.10. For the 3 language task on the DBIL data, the MODGDF (MGD) recognition

Table 6.10: Recognition performance of combined features for language identification.

Task	Feature	Database	Recogn.	Increase in Recogn.
Language Identification	MGD	DBIL 3 Language task	**96%**	-
	MFC		95%	-
	LFC		92%	-
	$\{MGD+MFC\}_{bm}$		98%	**2%**
	$\{MGD+MFC\}_{am}$		97%	**1%**
	$\{MFC+LFC\}_{bm}$		96%	**1%**
	$\{MFC+LFC\}_{am}$		95.5%	**0.5%**
Language Identification	MGD	OGI_MLTS 11 Language task	**53%**	-
	MFC		50%	-
	LFC		47%	-
	$\{MGD+MFC\}_{bm}$		58%	**5%**
	$\{MGD+MFC\}_{am}$		57%	**4%**
	$\{MFC+LFC\}_{bm}$		51%	**1%**
	$\{MFC+LFC\}_{am}$		50.5%	**0.5%**

performance was at 96%, MFCC (MFC) at 95%, $(\{\mathbf{MGD}+\mathbf{MFC}\}_{\mathbf{bm}})$ at 98%, and $(\{\mathbf{MGD}+\mathbf{MFC}\}_{\mathbf{am}})$ at 97% for this task. The best increase due to feature combination was 2%. For the 11 language task on the OGI_MLTS data, the MODGDF (MGD) recognition performance was at 53%, MFCC (MFC) at 50%, $(\{\mathbf{MGD}+\mathbf{MFC}\}_{\mathbf{bm}})$ at 58%, and $(\{\mathbf{MGD}+\mathbf{MFC}\}_{\mathbf{am}})$ at 57% for this task. The best increase due to feature combination was 5%. To enable a fair comparison the MFCC was also combined with LPCC and LFCC. While combining the MODGDF with the LPCC, brought down recognition performance, the combined pair of MODGDF and LFCC gave a small improvement in recognition performance. The results for combining the MODGDF with the LFCC are also tabulated in Table 6.10.

6.9.6 Experimental Results of Combined Features for Syllable Based Continuous Speech Recognition

In this Section, the experimental results for syllable based continuous speech recognition, on the DBIL database (Tamil and Telugu broadcast news) are discussed. The experimental setup is similar to that described in Section 6.6 for syllable recognition. The results of the MODGDF, MFCC and combined features on the DBIL database for both Tamil and Telugu are listed in Table 6.11. For Telugu broadcast news data, the MODGDF (MGD) recognition performance was at 38.2%, MFCC (MFC) at 38.6%, $(\{\mathbf{MGD}+\mathbf{MFC}\}_{\mathbf{bm}})$ at 50.6%, and $(\{\mathbf{MGD}+\mathbf{MFC}\}_{\mathbf{am}})$ at 44.6% for this task. The best increase due to feature combination was 12%. For Tamil broadcast news data, the MODGDF (MGD) recognition performance was at 36.7%, MFCC (MFC) at 37.1%, $(\{\mathbf{MGD}+\mathbf{MFC}\}_{\mathbf{bm}})$ at 48.2%, and $(\{\mathbf{MGD}+\mathbf{MFC}\}_{\mathbf{am}})$ at 42.2% for this task. The best increase due to feature combination was 11%. To enable a fair comparison the MFCC was also combined with LPCC and LFCC. While combining the MODGDF with the LPCC, brought down recognition performance, the combined pair of MODGDF and LFCC gave a small improvement in recognition performance. The results for combining the MODGDF with the LFCC are also tabulated in Table 6.11.

6.10 DISCUSSION

In this Chapter, the MODGDF is used for three speech recognition tasks syllable, speaker, and language recognition. The results suggest that the MODGDF can be used in practice across all speech recognition tasks. The significance of joint features derived by combining the short-time power and phase spectra is also investigated in this Chapter. It is illustrated that joint cepstral features derived from the modified group delay function and MFCC essentially capture complete formant information in

Table 6.11: Recognition performance of combined features for syllable recognition.

Task	**Feature**	**Database**	**Recogn.**	**Increase in Recogn.**
Syllable Recognition	MGD	DBIL Telugu broadcast news	**38.2%**	-
	MFC		38.6%	-
	LFC		32.6%	**-**
	$\{MGD + MFC\}_{bm}$		50.6%	**12%**
	$\{MGD + MFC\}_{am}$		44.6%	**6%**
	$\{MFC + LFC\}_{bm}$		41.6%	**3%**
	$\{MFC + LFC\}_{am}$		40.6%	**2%**
Syllable Recognition	MGD	DBIL Tamil broadcast news	**36.7%**	-
	MFC		37.1%	-
	LFC		31.2%	**-**
	$\{MGD + MFC\}_{bm}$		48.2%	**11%**
	$\{MGD + MFC\}_{am}$		42.2%	**5%**
	$\{MFC + LFC\}_{bm}$		39.2%	**2%**
	$\{MFC + LFC\}_{am}$		38.2%	**1%**

the speech signal. The advantage of using joint features for noisy data and related robustness issues are discussed. The joint features are also used for all the four speech recognition tasks phoneme, syllable, speaker, and language recognition. The results of the performance evaluation indicate that joint features improve recognition performance up to 12%, for feature combination before the acoustic model, and upto 6%, for likelihood combination, after the acoustic model. The results also indicate that the MODGDF complements the features derived from the short-time power spectra like the MFCC. Hence combining evidences derived from different feature streams and different systems, can be used to enhance the performance of speech recognition systems. The use of appropriate feature combinations before the model using mutual discriminant information together with a logical combination of the acoustic model outputs can further improve the recognition performance of the present day speech recognition systems.

CHAPTER 7

THE MULTI-STREAM AND SELECTIVE-BAND PARADIGM

Modern day automatic speech recognizers are highly sensitive to speaker and noise characteristics. Their performance degrades rapidly under mismatched training and testing environments. This is primarily because the knowledge gained from the training set does not apply well to the test data. The inability of single feature streams to capture time correlations and dynamics of the speech signal is another reason for poor recognition performance. One solution to the problem is the use of more appropriate feature extraction techniques. The use of new paradigms for automatic speech recognition based on the multi-stream and multi-band processing in tandem with newer feature extraction techniques is perhaps a promising way of achieving improved speech recognition performance under a rapidly changing application environment. Multi-stream processing considers the entire frequency band several times using a different processing strategy each time. Multi-band processing which belongs to the generic class of multi-stream processing techniques divides the entire frequency band into a set of frequency sub-bands and processes each band separately as an individual feature stream. The multi-stream and multi-band approaches alleviate the redundancy in the speech signal to some extent by either processing different parts of the signal separately or by processing the whole signal in different ways. Recombination at feature level is relevant when the multiple feature streams are correlated as it can model the dependencies in a better fashion. Feature streams that are uncorrelated are best modeled by multi-band approach. The multi-stream and multi-band approach is also used by the human perception system [6].

7.1 PSYCHO-ACOUSTIC MOTIVATION TO MULTI-STREAM AND MULTI-BAND SPEECH PROCESSING

Several psycho-acoustic studies have been conducted for the use of multiple feature streams in ASR

- Arai and Greenberg [59], showed that the robustness of spontaneous speech in terms of linguistic information lies across a wide span of time scales and frequency regions.
- Ghitza [60], investigated the use of perceptually related integration rules by humans. He concluded that humans use different frequency sub-bands and time scales to perceive short and long term information in the speech signal.
- Yang et al [61], describe the auditory cortex as a collection of sequential representation of the acoustic spectro-temporal information at different scales.
- Greenberg [62], showed the importance of longer time scale information to account for pronunciation variations.

7.2 REVIEW OF MULTI-STREAM SPEECH PROCESSING

Considerable amount of work has been carried within the framework of multi-stream and multi band processing. Some of the multi-stream approaches to speech recognition are enumerated below

- Antoniou and Reynolds [63, 64], used diverse feature streams like the MFCC, PLP, and LPC to derive an acoustic model for phone recognition. They showed that combination of several different front ends was better than the same artificial neural network trained on the same feature.
- Sharma et al [65], Kingsbury and Greenberg [66], Hermansky and Sharma [67], have used the concept of recombination on non-linearly transformed feature streams. All the experiments have shown that combining different non-linearly

transformed feature sets based on spectro-temporal processing gives a significant improvement in performance.

- Janin et al [53], and Cook et al [68], have investigated the use of averaging logarithmic probabilities from different feature streams within the HMM/MLP framework. They concluded that recombination of MLPs trained on multiple streams of features is more efficient than MLPs trained with different starting points.
- Ellis [55], compared recombination at both feature and probability level using PLP and MSG features. He showed that both combinations are equally efficient under various conditions.
- Shire [69], investigated the use of multiple front end acoustic modeling (PLP and MSG features) and recombining probabilities at frame level. He concluded that using multiple streams of features in the same training environment gave significant gains in recognition performance when the individual stream probabilities were combined.
- Wu et al [57], worked on combining features based on syllable and phone time scales at frame, syllable, and utterance level and concluded that recombination at frame level resulted in as good a performance as the recombination at syllable level.
- Kirchoff [70], and Kirchoff and Bilmes [71], tried various ways of combining acoustic and articulatory feature streams within the HMM/MLP framework.
- Christensen et al [72], and Meinedo and Neto [73], have investigated the use of diverse feature streams (PLP, RASTA-PLP, and MFCC) in the multi-stream and multi-band framework. It was concluded here that enhanced performance could be achieved with more diverse feature streams than homogeneous feature streams.
- Heckmann et al [74], have used a combination of both audio and visual features streams.

7.3 REVIEW OF MULTI-BAND SPEECH PROCESSING

Several investigations have been carried out on the multi-band paradigm which belongs to the generic multi-stream approach.

- Fletcher's work [6], was the first to identify that perception in humans involved sub-band processing.
- Nikki Mirghafori [75], developed specialized phone classes for each sub-band. His work concentrated on four sub-bands in an HMM framework. The best recognition was obtained when a PLP based sub-band system was combined with a RASTA-PLP full band system.
- Christophe Cerisara [76], implemented a four band system with second order HMMs. He combined individual sub-band likelihoods separately and also with the full band likelihood to achieve improved speech recognition performance.
- Stephane Dupont [77], investigated the use of several feature streams and several combination criteria within the multi-stream and multi-band framework.
- Hermansky et al [78], investigated multi-band systems consisting of 2 to 7 bands using different recombination strategies and concluded that sub-band systems with 2 bands reduce the WER by half.

7.4 THE MULTI-STREAM AND MULTI-BAND PARADIGM

Multi-stream (MS) speech recognition is based on the observation that different feature streams derived from the speech signal lead to different kinds of errors and on recombination some of these errors cancel out each other. In the multi-band (MB) approach several feature streams are processed in parallel, but these feature streams are all similar representations derived from different spectral regions of the speech signal.

7.4.1 The Multi-stream Paradigm

This paradigm looks at various ways in which speech recognition can be performed using different knowledge sources. The knowledge acquired from different representations of speech and different structures and type of classifiers are combined in this framework. This study is restricted to the combination of feature streams extracted with different techniques. The hypothesis is that the errors due to different representations of speech are complementary. The block diagram of the MS paradigm of speech recognition used in this work is given in Figure 7.1.

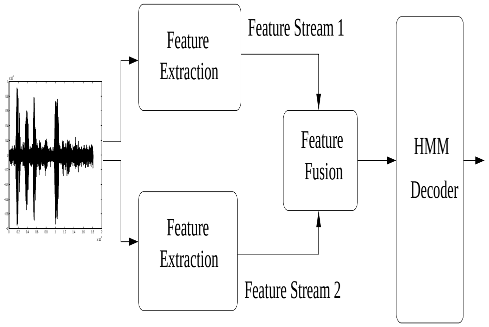

Fig. 7.1: Block diagram of the MS paradigm used in speech recognition.

7.4.2 The Multi-band Paradigm

The conventional approach to multi-band processing is to assume that individual sub-bands are independent of each other and extract features from each sub-band. The feature vectors hence extracted are called the feature sub-vectors. Recombination of these feature sub-vectors can be done in two ways. The feature sub-vectors can be concatenated into a single feature vector which is often called feature combination. Each of these feature sub-vectors can be processed independently and the posteriors can be combined by likelihood combination. Hence the two main approaches within the MB paradigm are the feature combination approach and the likelihood combination approach. In this study the feature combination approach is followed within the MB paradigm for speech recognition. The block diagram for such an approach is shown in Figure 7.2.

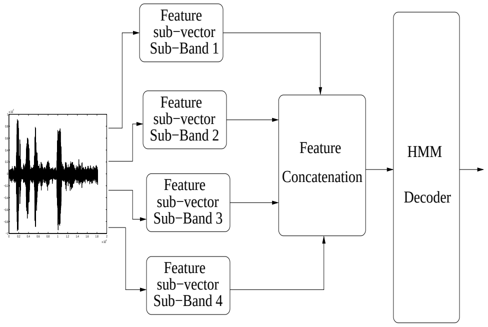
Feature sub−vector Sub−Band 1
Feature sub−vector Sub−Band 2
Feature sub−vector Sub−Band 3
Feature sub−vector Sub−Band 4
Feature Concatenation
HMM Decoder

Fig. 7.2: Block diagram of the MB paradigm used in speech recognition.

7.4.3 Mathematical Interpretation of the Product of Errors Rule for the MS-MB Formulation

Fletcher suggested in [6], that human beings decode linguistic message in the speech signal by processing narrow frequency bands independently of each other. Jont Allen [79], emphasized that combinations from these sub-bands is done at some intermediate level such that the product of the error rates in each sub-band is equal to the global error rate. Let x_1 and x_2, be two different (statistically independent) feature streams which make up the joint feature stream x=[x_1, x_2]. Let the error rate due to each individual feature stream be be $e(c_j|x_1)$ and $e(c_j|x_2)$ respectively, for any particular class c_j. Using the product of errors rule, the global error rate for the joint features stream x=[x_1, x_2], due to two individual feature streams x_1 and x_2, is therefore given by

$$e(c_j|x_1, x_2) = e(c_j|x_1)e(c_j|x_2) \tag{7.1}$$

From pattern recognition theory the probability of error P(e), for the observation vector x being misclassified as not belonging to the correct class is given by $1-P(c_j|x)$, where $P(c_j|x)$ is the probability of the observation vector x being classified as belonging to class j. Substituting this knowledge in Equation 7.1, we have

$$e(c_j|x_1, x_2) = (1 - P(c_j|x_1))(1 - P(c_j|x_2)) \tag{7.2}$$

Simplifying Equation 7.2, we have

$$e(c_j|x_1, x_2) = 1 - \sum_{i=1}^{2} P(c_j|x_i) + \prod_{i=1}^{2} P(c_j|x_i) \tag{7.3}$$

where $P(c_j|x_i)$ is the aposteriori probability of the i^{th} feature stream belonging to class j. But we do know that

$$P(c_j|x_1, x_2) = 1 - e(c_j|x_1, x_2) \tag{7.4}$$

Hence Equation 7.3, can now be rewritten using Equation 7.4 as

$$P(c_j|x_1, x_2) = \sum_{i=1}^{2} P(c_j|x_i) - \prod_{i=1}^{2} P(c_j|x_i) \tag{7.5}$$

From Equation 7.5, it is clear that $P(c_j|x_1, x_2)$ the probability of correct classification using a joint feature stream consisting of two feature streams x_1 and x_2 is equal to the diffrence of the probabilities $P(c_j|x_1)$ and $P(c_j|x_2)$ of correct classification due to individual feature streams. This is similar to the decision making mechanism in the human perception system as in [6].

7.5 THE MULTI-STREAM AND SELECTIVE-BAND PARADIGM

The two successful approaches to the MS-MB paradigm are the partial combination (PC) and the full combination (FC) approach [80]. In this Section, the single-stream full-band, multi-stream full-band, single-stream multi-band approaches to automatic speech recognition are discussed first. Within this framework a new approach called the multi-stream and and selective-band approach (MS-SB) is also proposed for automatic speech recognition.

7.5.1 The single-stream (SS-FB) paradigm

The single-stream (SS-FB) paradigm involves computing single stream features from the speech signal. The full frequency band of the speech signal is considered in this approach. The block diagram of this approach is given in Figure 7.3. The MODGDF and the MFCC are extracted from the full frequency band of the signal. One of the features is used in the succeeding steps of a speech recognition system.

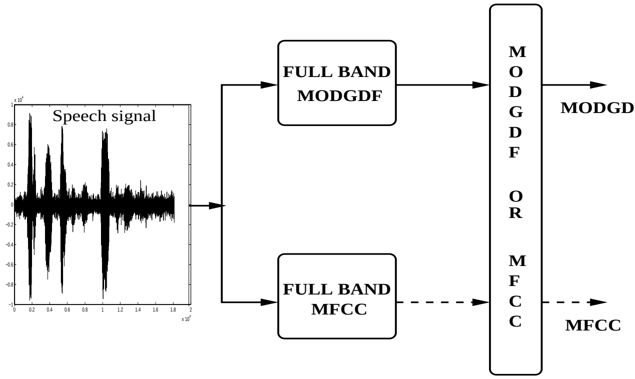

Fig. 7.3: Block diagram of the single-stream (SS-FB) paradigm

7.5.2 The multi-stream and full-band (MS-FB) paradigm

This paradigm looks at various ways in which speech recognition can be performed using different knowledge sources. The knowledge acquired from different representations of speech and different structures and type of classifiers are combined in this framework. In this paper we restrict ourselves to the combination of feature streams extracted with different techniques. The hypothesis is that the errors due to different representations of speech are complementary. The block diagram of the MS paradigm of speech recognition used in this work is given in Figure 7.4. The MODGDF and the MFCC are extracted from the full frequency band of the speech signal as shown in Figure 7.4. The two features are then combined using the early fusion and the joint

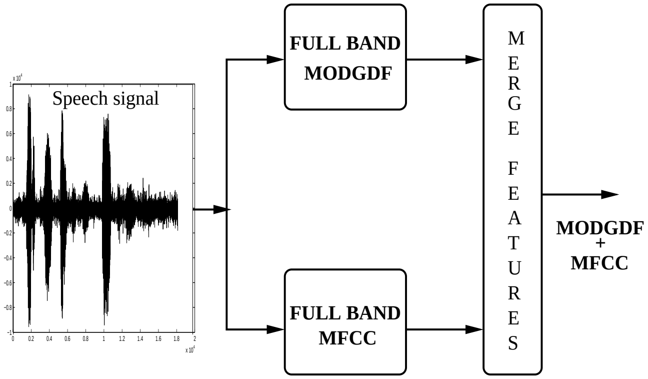

Fig. 7.4: Block diagram of the MS-FB paradigm

feature stream is used in the speech recognition process.

7.5.3 The single-stream and multi-band (SS-MB) paradigm

The conventional approach to multi-band processing is to assume that individual sub-bands are independent of each other and extract features from each sub-band. The features hence extracted are called the feature sub-vectors. Recombination of these feature sub-vectors can be done in two ways. The feature sub-vectors can be concatenated into a single feature vector which is often called feature combination. Each of these feature sub-vectors can be processed independently and the posteriors can be combined by likelihood combination. Hence the two main approaches within MB paradigm are the feature combination approach and the likelihood combination approach. In this paper we follow the feature combination approach within the MB paradigm for speech recognition. In the SS-MB approach the speech signal is passed through two triangular filters. The first filter passes the frequency band between 0 KHz to ($F_s/4$ + 1) KHz, while the second filter passes the frequency band between ($F_s/4$ - 1) KHz to $F_s/2$ KHz. F_s is the sampling rate. The response of the triangular filters used in this approach is shown in Figure 7.5. Note that the overlap is always equal to 2 KHz. The frequency band between 0 KHz to ($F_s/4$ + 1) KHz is called sub-band 1 and the frequency band between ($F_s/4$ - 1) KHz to $F_s/2$ KHz is called sub-band 2. The block diagram for such an approach is shown in Figure 7.6. The MODGDF is extracted from both the sub-bands 1 and 2. The features hence extracted are combined using early fusion to generate a joint feature stream MODGDF (sb1+sb2) as shown in Figure 7.6. Similarly

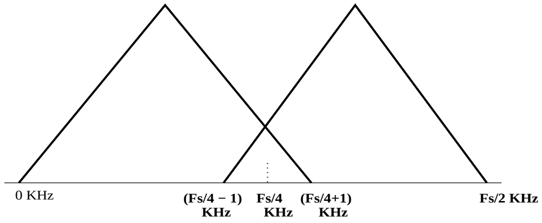

Fig. 7.5: Response of the two triangular filters used to derive the sub-bands in the SS-MB approach

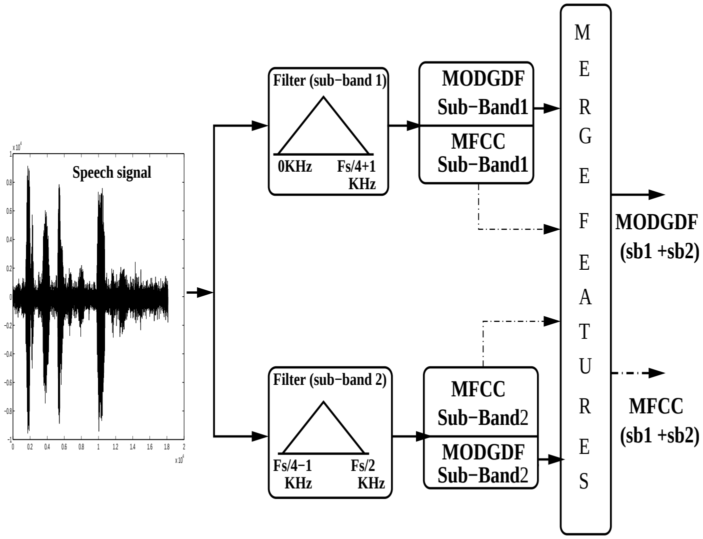

Fig. 7.6: Block diagram of the SS-MB paradigm the MFCC are derived from both the sub-bands 1 and 2 and concatenated to generate a joint feature stream MFCC (sb1+sb2) as shown in Figure 7.6. Note that the dotted lines indicate combination of the MFCC while the thick lines indicate combining the MODGDF. Both the multi-band MODGDF and the multi-band MFCC are then used individually in the speech recognition process.

7.5.4 The multi-stream and selective band (MS-SB) paradigm

The two approaches to the MS-MB paradigm are the partial combination (PC) and the full combination (FC) approach. Within this framework we propose a paradigm which is called the multi-stream and selective-band approach (MS-SB) for syllable based continuous speech recognition.

7.5.4.1 Analytical Formulation of the MS-SB approach

In the the MS-SB approach the speech signal is passed through two triangular filters. The first filter passes the frequency band between 0 KHz to (F_s/4 + 1) KHz, while the second filter passes the frequency band between (F_s/4 - 1) KHz to F_s/2 KHz. F_s is the sampling rate. The response of the triangular filters used in this approach is the same as shown in Figure 7.5. Note that the overlap is always equal to 2 KHz. Several approaches to multi-band processing pass the filter bank energies through a set of two filters. But in this approach we divide the signal into two frequency sub-bands and extract features from each sub-band. The implementation block diagram of the MS-SB approach is given in Figure 7.7. The feature merger is carried out in three different ways as shown in Figure 7.7.

1. The MODGDF derived from the low pass filtered sub-band (sub-band 1) and the MFCC from the low pass filtered sub-band (sub-band 1) are fused at feature merger stage. This combination is indicated by a thick line in Figure 7.7.
2. The MODGDF derived from the high pass filtered sub-band (sub-band 2) and the MFCC from the low pass filtered sub-band (sub-band 1) are fused at feature merger stage. This combination is indicated by a dashed line in Figure 7.7.

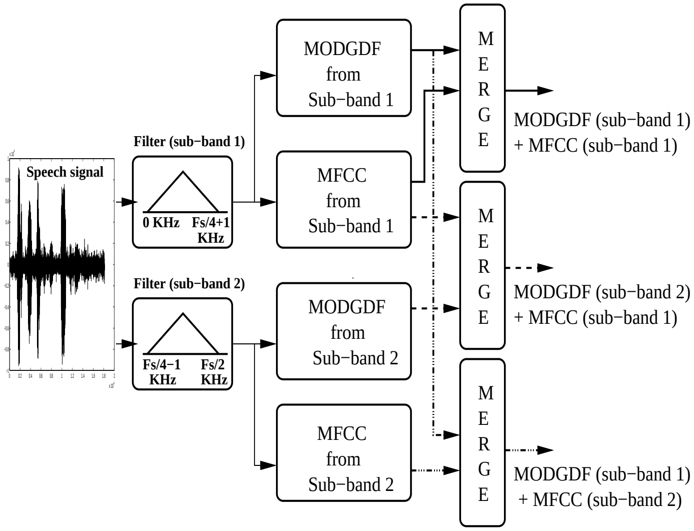

Fig. 7.7: Block diagram of the MS-SB paradigm

3. The MODGDF derived from the low pass filtered sub-band (sub-band 1) and the MFCC from the high pass filtered sub-band (sub-band 2) are fused at feature merger stage. This combination is indicated by a dash-dot line in Figure 7.7.

The three joint features thus derived using the MS-SB approach are used in the speech recognition process.

7.5.4.2 Extension of the MS-SB approach to the multi-stream-full-and-selective band (MS-FASB) approach

In the MS-SB paradigm discussed in Section 7.5.4, feature merger is carried out only on multiple streams of features across selective sub-bands. In an extension to this approach we combine features derived from the full frequency band along with the MS-SB derived features. This approach is called the multi-stream-full-and-selective band MS-FASB approach. The response of the triangular filters used for deriving sub-bands 1 and 2 are similar to that shown in Figure 7.5. The actual implementation block diagram of the MS-SB approach is given in Figure 7.8. In the MS-FASB paradigm the

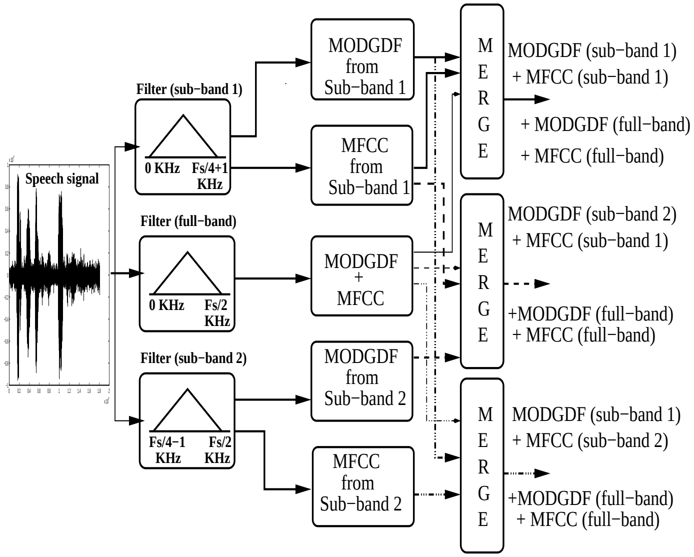
Filter (sub−band 1)
0 KHz
Fs/4+1 KHz
Speech signal
Filter (full-band)
0 KHz
Fs/2 KHz
Filter (sub−band 2)
Fs/4−1 KHz
Fs/2 KHz
MODGDF from Sub−band 1
MFCC from Sub−band 1
MODGDF + MFCC
MODGDF from Sub−band 2
MFCC from Sub−band 2
MERGE
MODGDF (sub−band 1) + MFCC (sub−band 1) + MODGDF (full−band) + MFCC (full−band)
MERGE
MODGDF (sub−band 2) + MFCC (sub−band 1) +MODGDF (full−band) + MFCC (full−band)
MERGE
MODGDF (sub−band 1) + MFCC (sub−band 2) +MODGDF (full−band) + MFCC (full−band)

Fig. 7.8: Block diagram of the MS-FASB paradigm

three types of joint features derived using the MS-SB approach are combined with the joint features (MODGDF + MFCC) extracted from the full band using early fusion. Hence three joint features are generated using the MS-FASB approach. These three features are then used in the speech recognition process.

7.6 PERFORMANCE EVALUATION

The SS-FB, SS-MB, MS-FB, MS-SB and MS-FASB approaches are used for syllable based continuous speech recognition on the DBIL [5] for two Indian languages Tamil and Telugu. The experimental conditions are very similar to that described in Chapter 6.

7.6.1 Databases and Baseline System used in the Study

The Database for Indian languages (DBIL) [5], used in this work consists of data from two Indian languages Tamil and Telugu. The baseline system uses HMM's (5 state and 3 mixture) as discussed in Chapter 6, for syllable based continuous speech recognition. The description of the databases is already given in Chapter 6.

7.6.2 Recognition Accuracy Evaluation Methodology

The baseline syllable recognition accuracy results using the segment based approach to recognition of continuous speech (discussed in Chapter 6), for two news bulletins each of duration fifteen minutes, comprising 9400 syllables for Tamil and Telugu using the SS-FB, SS-MB, MS-SB, and MS-FASB approaches are listed in the succeeding Sections. The baseline syllable recognition accuracy results using a conventional HMM (HTK) based recognition system are also listed. For both the baseline recognition system which uses a segment based approach to recognition of continuous speech and the conventional HMM (HTK) based recognition system, the following metric is used in computing the baseline syllable recognition accuracy.

- The baseline syllable recognition accuracy (SRA) is given by $\frac{(M-I)}{N}$ 100 % where M = Number of correctly recognized syllables, I is the number of insertions and N is the total number of syllables in the sentence that is recognized.
- Note that the SRA is raw syllable recognition accuracy without using any language models. The syllable recognition accuracy quoted here is similar to the raw baseline phoneme recognition accuracy in conventional phoneme based speech recognition systems without using language models.

7.6.3 Experimental Results for the Single-Stream Full-Band and Single-Stream Multi-Band Approach

The experimental results for the full-band single stream (SS-FB) and single-stream multi-band (SS-MB) approaches are listed in Table 7.1. It is significant to note that multi-band recombination on single feature streams gives an improvement of 2-3% in SRA. The multi-band recombination on the MFCC gives an improvement of 1.9% in SRA. The improvement is significant (2.4%) for a multi-band recombination on the MODGDF when compared to that of the MFCC (2%). The conventional HMM (HTK) based syllable recognition accuracy (HTKSRA) is also listed in Table 7.1. It is significant to note that the segment based recognition approach gives baseline syllable recognition accuracies greater than the conventional HMM (HTK) based recognition system.

Table 7.1: Baseline syllable recognition accuracy (SRA) and the conventional HMM (HTK) based syllable recognition accuracy (HTKSRA) of the MFC: MFCC, MGD: MODGDF and joint features for the DBIL data. (*sb*1 is the lower sub-band and *sb*2 is the upper sub-band).

Feature	Database	Approach	% SRA	% HTKSRA
MFC_{fb}	TAMIL	SS-FB	37.1	35.1
MGD_{fb}	TAMIL	SS-FB	36.7	34.2
$MFC_{sb1+sb2}$	TAMIL	SS-MB	39.1	37.0
$MGD_{sb1+sb2}$	TAMIL	SS-MB	38.1	36.2
MFC_{fb}	TELUGU	SS-FB	38.6	36.2
MGD_{fb}	TELUGU	SS-FB	38.2	35.9
$MFC_{sb1+sb2}$	TELUGU	SS-MB	41.6	39.0
$MGD_{sb1+sb2}$	TELUGU	SS-MB	40.6	38.0

7.6.4 Experimental Results for the Multi-Stream Full-Band and Multi-Stream Selective-Band Approach

In Table 7.2, the baseline SRA results of the multi-stream full-band (MS-FB) and multi-stream selective-band (MS-SB) approaches are listed. A significant improvement of 11% is achieved when compared to the MS-FB approach, while the SRA goes up by another 2-3% on using the MS-SB approach. Although all combinations between sub-bands 1 and 2 are tried, the best SRA is realized when the MODGDF extracted from the sub-band 1 (lower sub-band) is combined with the MFCC extracted from the sub-band 1 (lower sub-band). These results emphasize the fact that the lower formants are useful for speech recognition. The conventional HMM (HTK) based syllable recognition accuracy (HTKSRA) is also listed in Table 7.2.

Table 7.2: Baseline syllable recognition accuracy (SRA) and the conventional HMM (HTK) based syllable recognition accuracy (HTKSRA) of the joint features using the MS-FB and MS-SB approaches for the DBIL data. (*sb*1 is the lower sub-band and *sb*2 is the upper sub-band).

Feature	Database	Approach	% SRA	% HTKSRA
$MGD_{fb} + MFC_{fb}$	TAMIL	MS-FB	48.2	46.0
$MGD_{sb2} + MFC_{sb1}$	TAMIL	MS-SB	50.0	47.8
$MGD_{sb1} + MFC_{sb2}$	TAMIL	MS-SB	49.5	47.4
$MGD_{sb1} + MFC_{sb1}$	TAMIL	MS-SB	51.9	49.8
$MGD_{fb} + MFC_{fb}$	TELUGU	MS-FB	50.6	48.6
$MGD_{sb2} + MFC_{sb1}$	TELUGU	MS-SB	51.6	49.5
$MGD_{sb1} + MFC_{sb2}$	TELUGU	MS-SB	51	48.9
$MGD_{sb1} + MFC_{sb1}$	TELUGU	MS-SB	53.6	51.4

7.6.5 Experimental Results for the Multi-Stream Full and Selective-Band Approach

The Multi-stream full and selective-band (MS-FASB) approach involves combining features derived from the entire frequency band and features derived from the MS-SB approach. The SRA results for syllable recognition using the MS-FASB approach are listed in Table 7.3. The improvements noticed are less and in some cases there is no improvement in recognition performance. The conventional HMM (HTK) based syllable recognition accuracy (HTKSRA) is also listed in Table 7.3.

Table 7.3: Baseline syllable recognition accuracy (SRA) and the conventional HMM (HTK) based syllable recognition accuracy (HTKSRA) of the joint features using the MS-FASB approach for the DBIL data. ($sb1$ is the lower sub-band and $sb2$ is the upper sub-band).

Feature	Database	% SRA	% HTKSRA
$MGD_{fb} + MFC_{fb} + MGD_{sb2} + MFC_{sb1}$	TAMIL	50.5	48.3
$MGD_{fb} + MFC_{fb} + MGD_{sb1} + MFC_{sb2}$	TAMIL	50.0	48.3
$MGD_{fb} + MFC_{fb} + MGD_{sb1} + MFC_{sb1}$	TAMIL	52.0	49.8
$MGD_{fb} + MFC_{fb} + MGD_{sb2} + MFC_{sb1}$	TELUGU	52.0	49.8
$MGD_{fb} + MFC_{fb} + MGD_{sb1} + MFC_{sb2}$	TELUGU	51.5	49.5
$MGD_{fb} + MFC_{fb} + MGD_{sb1} + MFC_{sb1}$	TELUGU	53.7	51.0

7.7 DISCUSSION

In this Chapter, two new paradigms are proposed for improved speech recognition. The MS-SB and MS-FA-SB paradigms are proposed within the framework of the multi-stream (MS) and multi-band (MB) approaches for automatic speech recognition. Indeed both the MS and MB approaches give significant impovements in recognition performance. But the knowledge of which bands are important for different feature streams is quite crucial for improved speech recognition performance. An effort has been made in this direction by combining different features derived from different sub-bands for continuous speech recognition. The MODGDF and its significance in speech processing has been discussed earlier in the thesis. But the idea of combining the Fourier transform magnitude and phase spectra for representing speech via the group delay domain and MFCC within the MS and MB framework is presented in this

Chapter. The joint features derived using the MS-FB, MB-SS, MS-FB, MS-SB, and MS-FA-SB approaches are used for syllable based continuous speech recognition for two Indian languages Tamil and Telugu. The results of the performance evaluation indicate that joint features improve recognition performance by 2-3% for the MB-SS approach, up to 12% for the MS-FB approach, and upto 13% for the MS-SB approach, when compared to the full-band single-stream approach. The MS-FA-SB approach gave little improvement over the MS-SB approach. The results from this study indicates that combining evidences derived from diverse feature streams derived from multiple frequency bands and different recombination schemes does enhance recognition performance of speech recognition systems. The use of appropriate feature combinations from selective frequency bands using mutual discriminant information together with a full combination approach can further improve speech recognition performance.

CHAPTER 8

SUMMARY AND CONCLUSIONS

8.1 SUMMARY

This thesis describes methods for formulating a new spectrum called the modified group delay spectrum and extracting speech features from it. The idea of utilizing the Fourier transform phase for representing speech via the group delay domain is presented in this thesis. The group delay function becomes spiky due to roots that are very close to the unit circle in the *z*-plane and also due to pitch periodicity effects. The computation of the group delay function is modified to alleviate these effects. Cepstral features are derived from the modified group delay function, which are highly decorrelated and essentially robust to the channel and white noise.

It is significant to note that the MODGDF exhibits compact cluster structures and good class separability in the lower dimensional feature space. From the results presented in this work, it is clear that the MODGDF captures the dynamic information of the speech signal. The MODGDF is used for three speech recognition tasks syllable, speaker, and language recognition and the results suggest that the MODGDF is a feature that can be used in practice across all speech recognition tasks. The good pairwise separability exhibited by the MODGDF indicates that it could be the feature of choice in speech processing tasks where a few classes have to be recognized. The MODGDF as the front end and a suitable pairwise classifier would work extremely well in these applications.

This thesis also analyzes the significance of joint features derived by combining short time magnitude and phase spectra. It is illustrated that combining cepstral features derived from the modified group delay function and MFCC essentially capture complete formant information in the speech signal. It is evident from the results presented in this work that the MODGDF complements the features derived from the Fourier transform magnitude like MFCC. The combined features are used for three speech recognition tasks syllable, speaker recognition, and language recognition. The results of the performance evaluation indicate that combining features improves recognition performance up to 12% for feature combination before the acoustic model and upto 6% for feature combination after the acoustic model. This clearly indicates that combining evidences derived from different feature streams and different systems, does enhance recognition performance of speech recognition systems. The use of appropriate feature combinations before the model using mutual discriminant information together with a logical combination of the classifier outputs can improve recognition performance of the present day, state of the art, speech recognition systems.

This thesis also proposes two new paradigms for improved speech recognition. The MS-SB and MS-FA-SB paradigms proposed in this work are derived from within the framework of the multi-stream and multi-band approaches used in speech recognition. Indeed the both the MS and MB approaches give significant improvements in recognition performance. But the knowledge of which bands are important for different feature streams for improving speech recognition performance is crucial. The idea of combining features derived from the Fourier transform magnitude and phase, within the MS and MB framework is presented. It is illustrated using formant reconstruction techniques that joint features derived from the modified group delay function and MFCC from the MS-SB approach essentially capture complete formant information in the speech signal in terms of spectral resolution and smoothness. The joint features derived using the MS-FB, MB-SS, MS-FB, MS-SB, and MS-FA-SB approaches are used for syllable based continuous speech recognition for two Indian languages Tamil and Telugu. The results of the performance evaluation indicate that joint features improve recognition performance by 3-4% for the FB-SS approach, up to 11% for the MS-FB approach, and upto 14% for the MS-SB approach. The use of appropriate feature combinations from selective frequency bands using mutual discriminant information together with a full combination approach can significantly improve speech recognition performance.

8.2 KEY IDEAS PRESENTED IN THE THESIS

- A new representation of speech called the modified group delay feature (MODGDF), which is based on the Fourier transform phase is proposed.
- The suitability of the MODGDF as a feature for speech recognition applications is studied extensively using a number of different criteria.
- The complementary nature of the MODGDF derived from the modified group delay function with respect to features derived from the Fourier transform magnitude spectra (MFCC) is illustrated with the help of extensive experiments.
- A new approach called the multi-stream and selective-band approach is proposed within the framework of the multi-stream and multi-band paradigm, for improved speech recognition performance.
- The significance of the MODGDF in the multi-stream and multi-band paradigm of automatic speech recognition is highlighted.

8.3 CRITICISMS

The following are the criticisms regarding the work presented herein.

1. The values of the two parameters α and γ used in the computation of the MODGDF have been optimized using a minimization of mean square error approach, and also using line search across several databases. But it is not very clear how these values can be optimized on unseen speech data in an adaptive manner.
2. The values of α and γ are optimized using an iterative approach. Hence the optimization can be called empirical rather than absolute.
3. The MODGDF outperforms the MFCC in terms of crucial feature evaluation criteria like discriminability and cumulative separability. But in terms of recognition, the performance of MODGDF is not that spectacular when compared to the MFCC.

4. Although the MODGDF works well across speech recognition applications, it is not very clear what is the complementary information of the speech signal that the MODGDF captures, when compared to that of the MFCC.

8.4 SCOPE FOR FUTURE WORK

- A closed form solution to estimate the values of the two free parameters α and γ used in the computation of the MODGDF needs to be worked out.
- There is need to investigate the possibility of α and γ being related to some measure of the speech signal.
- The delta and delta-delta computation from the MODGDF has to be refined to improve speech recognition performance.
- Since experiments on discriminability analysis suggest that the MODGDF linearly discriminates between a pair of classes, a new classifier that is designed to capture this information can further improve recognition performance.
- A measure of complementarity should be defined for multiple-feature streams prior to their fusion. The measure could be at the signal processing level or at the recognition level.
- Although it has been proved that the ear is non linearly sensitive to spectral magnitude, it needs to be found by perception experiments whether the same is true for spectral phase.
- The issue of frequency warping in the modified group delay domain needs to be addressed.

APPENDIX A

DISTANCE MEASURES IN SPEECH PATTERN CLASSIFICATION

The primary task of a speech recognition system is to discriminate between speech patterns. The speech signal is generally represented as a set of feature vectors distributed in a higher dimensional feature space. The classification of these speech feature vectors is a task of evaluating the difference between (or rather the similarity between) these patterns. The dissimilarity between speech patterns is generally expressed as distance measure. The distance measures are often computed over a higher dimensional feature space. These distance measures depend on the correlation and the weights of different dimensions of the feature vector.

A.1 SIMPLE DISTANCE MEASURES

The discrimination between any two templates (representing speech patterns) is directly proportional to the distance between the points corresponding to these templates in the higher dimensional feature space. The most common distance measures used in speech recognition are the Euclidean and the Mahalanobis distance.

- **The Euclidean distance (L^2-norm)** is computed as

$$E(x,y) = \sqrt{(x-y)^T(x-y)} \tag{A.1}$$

It can also be expressed as

$$E(x,y) = \sqrt{\sum_{i=1}^{N}(x_i - y_i)^2} \tag{A.2}$$

where x and y are any two points distributed in the N dimensional feature space.

- **The Mahalanobis Distance (covariance weighted distance)** is computed as

$$M(x,y) = \sqrt{(x-y)^T W^{-1}(x-y)} \tag{A.3}$$

where W is a positive definite (auto covariance) matrix which is used to weight the speech feature vectors in the N dimensional feature space.

A.2 PROBABILISTIC DISTANCE MEASURES

An efficient distance measure for discriminating speech patterns should capture both the intra-class and the inter-class variablity of speech patterns. Probabilistic distance measures are most suited for speech recognition applications where the high variability in speech has to be captured. Probabilistic techniques model these these variations as multivariate probability distribution functions (pdf's) over a higher dimensional feature space. These multivariate pdf's are used to identify a particular speech pattern by maximizing the likelihood of the speech pattern belonging to a particular class. For a univariate Gaussian case, a different Gaussian pdf can be associated with each acoustic class. The Gaussian pdf of a feature vector x for the i^{th} acoustic class is given in Equation A.4.

$$p_i(x) = \frac{1}{(2\pi)^{R/2}|\Sigma_i|^{1/2}} e^{-\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i)} \tag{A.4}$$

where μ_i is the state mean vector , Σ_i the state covariance matrix, and R the dimension of the feature vector. $(x - \mu_i)^T$ is the transpose of the matrix $(x - \mu_i)$. Σ_i^T is the transpose and Σ_i^{-1} is the inverse of the matrix Σ_i. The expected value of the feature vector x is given by the mean μ_i. For a multivariate Gaussian case, the probability that a feature vector being in any one of the I acoustic classes for a particular model model λ, is given by the mixture, of different gaussian pdfs

$$p_i(x|\lambda) = \sum_{i=1}^{I} w_i p_i(x) \tag{A.5}$$

where $p_i(x)$ are the mixture densities and w_i are the mixture weights. Note that sum of all the weights is equal to one as in Equation A.6.

$$\sum_{i=1}^{I} w_i = 1 \tag{A.6}$$

Using the Bayes rule for maximum likelihood estimation a test pattern x is declared as belonging to class 1 if

$$p_1(x) \geq p_i(x) \tag{A.7}$$

where $i = \{1,2,3, ... , \text{N}\}$ and N is the total number of classes.

A.3 CEPSTRAL DISTANCE MEASURES

Cepstral features are perhaps the most widely used features for speech recognition. Although a simple Euclidean distance measure can be used to compute the distance between cepstral vectors, a modified version of this distance measure which is called the RMS log spectral distance is used to evaluate the distance between cepstral vectors. This cepstral distance measure is given by

$$C^2(r,t) = \sum_{n=-\infty}^{\infty} (r_n - t_n)^2 \tag{A.8}$$

It can also be expressed as

$$C^2(r,t) = (2\pi)^{-1} \int_{\omega=-\pi}^{\pi} |logR(\omega) - logT(\omega)|^2 d\omega \tag{A.9}$$

where r_n and t_n are the cepstral feature vectors for the reference and the test speech patterns respectively.

APPENDIX B

VECTOR QUANTIZATION

Given $x = \{x_1, x_2, ..., x_d\} \ \epsilon \ R^d$, where x is a d-dimensional vector whose components $\{x_k, 1 \leq k \leq d\}$, are real valued, vector quantization maps the vector x to another discrete amplitude d dimensional vector z, such that $z = q(x)$, where $q()$ is the quantization operator. Generally z takes one of the finite set of values $Z = \{z_i, 1 \leq i \leq L\}$, where $z = \{z_1, z_2, z_3, ..., z_d\}$. The finite set Z, is referred to as the codebook and $\{z_i\}$, are the codewords, where L the size of the codebook. L is also called the number of levels in the codebook. Codebook design involves partitioning of the d-dimensional space of the observation vector x into L regions or cells $\{C_i, 1 \leq i \leq L\}$. Associated with each cell C_i is a vector z_i. The quantizer assigns the codeword z_i if x lies in cell C_i. This process of assignment is also called training the codebook. An example of partitioning the codebook is shown in Figure 8.4. The regions enclosed by lines within the rectangular box shown in Figure 8.4 are the cells C_i. All vectors lying inside cell C_i are quantized as z_i. The positions of the codewords corresponding to each cell

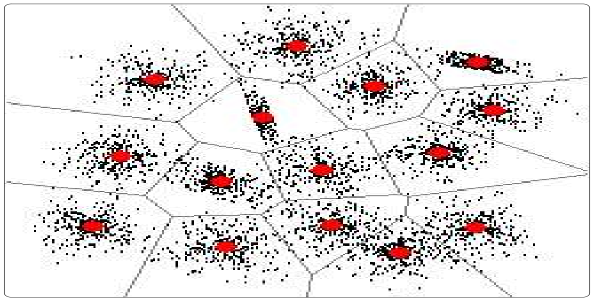

Fig. B.1: Partitioning of the 2 dimensional sample space into 15 cells using vector quantization.

is determined by minimizing the average distortion associated with each cell. If x is quantized to z, then the distortion measure between x and z can be computed in several ways. The most commonly used measure is the Euclidean distance measure. The Mahalanobis distance measure is also widely used to give unequal weights to certain contributions which make up the overall distortion. The overall average distortion is computed as

$D = E[d(x, z)]$

where $E[.]$ is the expectation of $d(x, z)$. Considering $d(x, z)$ to be the Mahalanobis distance as in Equation A.3, the overall average distortion equals

$$D = \sum_{i=1}^{L} D_i \tag{B.10}$$

where D_i denotes the average distortion in cell C_i. For a given speech pattern there is no optimal solution to minimize the overall average distortion. Iterative algorithms like the k-means algorithm [81] and the *LBG* algorithm [82] are generally used in this context.

THE K-MEANS ALGORITHM

The k-means algorithm [82] is described by the following steps

1. Derive an initial codebook using an adequate method (for eg. Random codebook method)
2. Classify each element of the training vector set $\{x_k\}$ into one of the clusters C_i, by selecting the codeword z_i which is closest to $\{x_k\}$. The closeness is generally measured by the Euclidean distance.
3. Update the codeword of every cluster by computing the centroid of the training vectors in each cluster.
4. If the decrease in the overall average distortion at this iteration is less than a particular threshold in comparison to the overall distortion in the previous iteration, Go to step 2.
   Else Stop.

The disadvantage of the k-means algorithm is that it does not converge to a global optimum. The *LBG* algorithm [82], is a modified version of the k-means algorithm, where the centroid is determined by iterative refinements over k-means clustering.

APPENDIX C

GAUSSIAN MIXTURE MODELING

The speech production mechanism is not deterministic. The high degree of variability in speech can be modelled by a multivariate Gaussian probability distribution function (pdf). For a univariate Gaussian case, a different Gaussian pdf can be associated with each acoustic class. The Gaussian pdf of a feature vector x for the i^{th} acoustic class is given in Equation C.11.

$$p_i(x) = \frac{1}{(2\pi)^{R/2}|\Sigma_i|^{1/2}} e^{-\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i)} \tag{C.11}$$

where μ_i is the state mean vector , Σ_i the state covariance matrix, and R the dimension of the feature vector. $(x - \mu_i)^T$ is the transpose of the matrix $(x - \mu_i)$. Σ_i^T is the transpose and Σ_i^{-1} is the inverse of the matrix Σ_i. The expected value of the feature vector x is given by the mean μ_i. For a multivariate Gaussian case, the probability that a feature vector being in any one of the I acoustic classes for a particular model model λ, is given by the mixture, of different gaussian pdfs

$$p_i(x|\lambda) = \sum_{i=1}^{I} w_i p_i(x) \tag{C.12}$$

where $p_i(x)$ are the mixture densities and w_i are the mixture weights. Note that sum of all the weights is equal to one as in Equation C.13.

$$\sum_{i=1}^{I} w_i = 1 \tag{C.13}$$

A model that consists of the component mixture densities, means, and co-variances derived from the feature vector, is called the Gaussian mixture model (GMM) of that acoustic class. The GMM λ corresponding to a particular acoustic class is given by Equation C.14.

$$\lambda = \{p_i, \mu_i, \Sigma_i\} \tag{C.14}$$

where p_i are the component mixture densities, μ_i the means, and Σ_i the co-variances.

C.1 TRAINING THE MODEL

The training procedure is similar to the procedure followed in vector quantization. Clusters are formed within the training data. Each cluster is then represented with multiple Gaussian pdf's. The union of many such Gaussian pdfs is a GMM. The most common approach to estimate the GMM parameters is the maximum likelihood estimation [83], where $p(X|\lambda)$ is maximized with respect to λ. $p(X|\lambda)$ is the conditional probability and vector $X = \{x_0, x_1, ..., x_{M-1}\}$ is the set of all feature vectors belonging to a particular acoustic class. Since there is no closed form solution to the maximum likelihood estimation, convergence is guaranteed only when large enough data is available. An iterative approach using expectation-maximization (EM) algorithm [82] is followed. The EM algorithm improves on the GMM parameter estimates by iteratively checking for the condition

$$p(X|\lambda^{k+1}) > p(X|\lambda^k) \tag{C.15}$$

wher k is the number of iterations.

C.2 DECISION MAKING IN THE SPEECH CONTEXT

Let the number of models representing different acoustic classes be S. Hence λ_j, where $j = \{1, 2, 3, ..., S\}$, is the set of GMMs under consideration. For each test utterance, feature vectors x_n at time n are extracted. The probability of each model given the feature vectors x_n is given by

$$P(\lambda_j|x_n) = \frac{p(x_n|\lambda_j)P(\lambda_j)}{P(x_n)} \tag{C.16}$$

Since $P(x_n)$ is a constant and $P(\lambda_j)$ the apriori probabilities are assumed to be equal, the problem is reduced to finding the λ_j that maximizes $p(x_n|\lambda_j)$. But $p(x_n|\lambda_j)$ is given by

$$P(x_n|\lambda_j) = p(\{x_0, x_1, ..., x_{M-1}\}|\lambda_j) \tag{C.17}$$

where M is the number of feature vectors for each frame of the speech signal belonging to a particular acoustic class. Assuming that each frame is statistically independent Equation C.17 can be written as

$$p(\{x_0, x_1, ..., x_{M-1}\}|\lambda_j) = \prod_{m=0}^{M-1} p(x_m|\lambda_j) \tag{C.18}$$

Applying logarithm on Equation C.18 and simplifying for S we have

$$S_r = \underset{1 \leq j \leq s}{max} \sum_{m=0}^{M-1} log[p(x_m|\lambda_j)] \tag{C.19}$$

where S_r is declared as the class to which the feature vectors belong. Note that $\{S_r, r = \{1, 2, 3, ..., S\}\}$ is the set of all acoustic classes. Gaussian mixture modeling techniques are widely used in speaker and language identification.

APPENDIX D

HIDDEN MARKOV MODELING

Template matching methods assume that reference patterns represent the true pattern and therefore the recognition system is conceptually simple. Hidden Markov model (HMM) is a statistical method of characterizing the spectral properties of a speech pattern. For modeling speech patterns, HMMs are suitable as speech can be characterized as a parametric random process. HMMs can handle the durational variations and can incorporate both the task syntax and semantics. HMMs are finite state diagrams, where the states are hidden.

Each transition in the state diagram of a HMM has a transition probability associated with it. These transition probabilities are denoted by matrix A. Here A is defined as, $A = a_{ij}$ where $a_{ij} = P(i_{t+1} = j|i_t = i)$, the probability of being in state j at time $t + 1$, given that we were in state i at time t. It is assumed that $a_{ij}'s$ are independent of time.

Each state is associated with a set of discrete symbols with an observation probability assigned to each symbol, or is associated with the set of continuous observations with a continuous observation probability density. These observation symbol probabilities are denoted by the parameter B. Here B is defined as, $B = b_j(k)$, $b_j(k) = P(v_k \; at \; t|i_t = j)$, the probability of observing the symbol v_k, given that we are in state j.

The initial state probability is denoted by the matrix π, where π is defined as
$\pi = \pi_i, \pi_i = P(i_1 = i)$, the probability of being in state i at $t = 1$.

Using the three parameters A, B and π, a HMM can be compactly denoted as $\lambda = \{A, \ B, \ \pi\}$.

An example for HMM with three states and three mixtures per state is shown in Figure D.2.

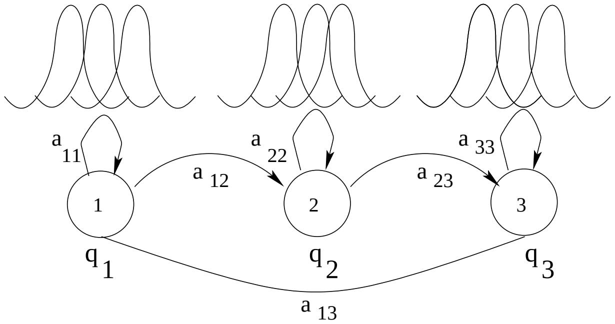

Fig. D.2: Topology of a 3 state left to right continuous HMM with three mixtures per state.

The three basic problems and their solutions in implementing HMMs [81] are

1. Evaluation (Scoring)
    - Problem: Given a model and a sequence of observations, how do we compute the probability that the observed sequence was produced by the model? In other words how well a given model matches a given observation sequence.
    - Solution: Forward method or Backward method
2. Finding the optimal state sequence (Testing)
    - Problem: To find the optimal state sequence associated with the given observation sequence.
    - Solution: Viterbi algorithm
3. Parameter estimation (Training)
    - Problem: To determine a method to adjust the model parameters so as to satisfy an optimization criterion.

    - Solution: Expectation maximization (EM) method.

The solution to problem 1 is useful in the recognition phase. In this phase, for the given parameter vector sequence (observation sequence) derived from the test speech utterance, the likelihood value of each HMM is computed using the *forward procedure* [84]. Here, one HMM corresponds to one speech unit. The symbol associated with the HMM, for which the likelihood is maximum, is identified as the recognized symbol corresponding to the input speech utterance.

Problem 2, is associated with training of the HMM for the given speech unit. Several examples of the same speech segment with different phonetic contexts are taken, and the parameters of the HMMs, λ, have been iteratively refined for maximum likelihood estimation, using the $Baum-Welch$ algorithm [85]. Parameter estimation of the HMM's, where each state is associated with mixtures of multi variate densities has been demonstrated in [83].

Problem 3 is associated with several applications, including the segmentation of the speech signal. The Viterbi algorithm [86, 87] is employed for solving the problem 3 as it is computationally efficient.

VITERBI ALGORITHM

Viterbi algorithm is one application of dynamic programming. Given an observation sequence derived from a speech utterance, the Viterbi algorithm finds the most likely state sequence and the likelihood value associated with this most likely sequence in a given HMM [18, 81]. Given a set of HMM's trained for predetermined set of speech units, Viterbi algorithm can also be used during the recognition phase to determine the HMM from the set of HMM's that matches best with a given observation sequence.

In the state sequence estimation problem, a set of T observations, $\mathbf{O} = \{\mathbf{o_1}, \mathbf{o_2}, \ldots,$ and an N state HMM, λ are given. The goal is to estimate the state sequence, $S = \{s(1), s(2), \ldots, s(T)\}$ which maximizes the likelihood $L(\mathbf{O}|\mathbf{S}, \lambda)$. Determining the most likely state sequence can be solved by using dynamic programming [86, 87]. Let $\phi_j(t)$ represent the probability of the most likely state sequence for observing vectors o_1 through o_t, while at state j, at time t, and $B_j(t)$ represents the state which gives this probability, then $\phi_j(t)$ and $B_j(t)$ can be expressed as

$$\phi_j(t) = max_i\{\phi_j(t-1).a_{ij}\}.b_j(o_t) \tag{D.20}$$

$$B_j(t) = arg(max_i\{\phi_j(t-1).a_{ij}\}.b_j(o_t)) \tag{D.21}$$

using initial conditions,

$$\phi_1(1) = 1 \tag{D.22}$$

$$B_1(1) = 0 \tag{D.23}$$

$$\phi_j(1) = a_{1j}.b_j(o_t) \quad for \quad 1 < j \leq N \tag{D.24}$$

$$B_j(1) = 1 \tag{D.25}$$

In Equation D.20, the probability $\phi_j(t)$ is computed using a recursive relation. Using $B_j(t)$ and assuming that the model must end in the final state at time T, $(s(T) = N)$, the sequence of states for the maximum likelihood path can be recovered recursively using the equation,

$$s(t-1) = B_{s(t)}(t). \tag{D.26}$$

In otherwords, starting with $s(T)$ known, Equation D.26 gives the maximum likelihood state at time $T-1$ (e.g $s(T-1) = B_{s(T)}(t) = B_N(t)$). The observation made in the Viterbi algorithm is that, for any state at time t, there is only one most likely path to that state. Therefore, if several paths converge to a particular state at time t, instead of recalculating all of them when calculating the transitions from this state to states at time $t+1$, one can discard the less likely paths, and only use the most likely one in the calculations. When this is applied to each time step, the number of calculations is reduced to $T.N^2$, which is much lesser than T^N.

REFERENCES

[1] **L. Liu, J. He and G. Palm**, (1997), Effects of phase on the perception of intervocalic stop consonants, *Speech Communication*, vol. 22, pp. 403–417.

[2] **A. V. Oppenheim and R. W. Schafer**, *Discrete-time signal processing.* Prentice Hall, 2000.

[3] **NTIS**, *The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus.* 1993.

[4] **C. Jankowski, A. Kalyanswamy, S. Basson, and J. Spitz**, (1990), *NTIMIT* : A Phonetically Balanced, Continuous Speech, Telephone Bandwidth Speech Database, *Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing*, April, pp. 109–112.

[5] *Database for Indian languages.* IIT Madras, Chennai, India: Speech and Vision Lab, 2001.

[6] **H. Fletcher**, *Speech and Hearing in Communication.* New York: Krieger.

[7] **D. Ellis**, *Feature stream combination before and/or after the acoustic model.* Berkeley: ICSI Technical Report, 2000.

[8] **H. A. Murthy**, *Algorithms for Processing Fourier Transform Phase of Signals.* PhD dissertation, Indian Institute of Technology, Department of Computer Science and Engg., Madras, India, December 1991.

[9] **H. Drucker**, (1968), Speech processing in a high ambient noise environment, *IEEE Trans. Electroacoustics*, vol. AU-16, pp. 165–168.

[10] **H. Helmholtz**, *On the Sensations of Tone.* New York: Dover, 1954.

[11] **K. K. Paliwal**, (2004), Usefulness of phase in speech processing, *Journal of Negative Results in Speech and Audio Sciences*, p. Article 2.

[12] **H. A. Murthy and B. Yegnanarayana**, (1991), Formant extraction from group delay function, *Speech Communication*, vol. 10, pp. 209–221.

[13] **B. Yegnanarayana and H. A. Murthy**, (1992), Significance of group delay functions in spectrum estimation, *IEEE Trans. Signal Processing*, vol. 40, pp. 2281–2289.

[14] **V. K. Prasad, T. Nagarajan, and H. A. Murthy**, (2004), Automatic segmentation of continuous speech using minimum phase group delay functions, *Speech Communication*, vol. 42, pp. 429–446.

[15] **V. K. Prasad**, *Segmentation and Recognition of Continuous Speech.* PhD dissertation, Indian Institute of Technology, Department of Computer Science and Engg., Madras, India, 2003.

[16] **S. S. Stevens and J. Volkman**, (1940), The relation of pitch to frequency, *American Journal of Psychology*, vol. 53(3), pp. 329–353.

[17] **P. Mermelstein and S. B. Davis**, (1980), Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences, *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 28, pp. 357–366.

[18] **J. W. Picone**, (1993), Signal modeling techniques in speech recognition, *Proc. IEEE*, vol. 81(9), pp. 1215–1247.

[19] **H. Sheikhzadeh and L. Deng**, (1994), Waveform based speech recognition using hidden filter models: Parameter selection and sensitivity to power normalization, *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 2(1), pp. 80–90.

[20] **O. Ghitza**, (1994), Auditory models and human performance in tasks related to speech coding and speech recognition, *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 2(1), pp. 115–132.

[21] **K. L. Payton**, (1988), Vowel processing by a model of the auditory periphery: A comparison to eight nerve responses, *J. Acoust. Soc. Amer.*, vol. 83, pp. 145–162.

[22] **R. F. Lyon**, (1982), A computational model of filtering, detection and compression in the cochlea, *Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing*, vol. 2(1), pp. 1282–1285.

[23] **S. Seneff**, (1988), A joint synchrony/mean rate model of auditory speech processing, *J. Phonetics*, vol. 16(1), pp. 55–76.

[24] **B. P. Bogert, M. J. R. Healy, and J. W. Tukey**, (1963), The quefrency analysis of time series for echoes: Cepstrum, pseudo-autocovariance, cross cepstrum, and saphe cracking, *Proceedings of the Symposium on Time Series Analysis*, John Wiley and Sons, New York, pp. 209–243.

[25] **A. M. Noll**, (1967), Cepstrum pitch determination, *J. Acoust. Soc. Amer.*, pp. 179–195.

[26] **L. R. Rabiner and B. H. Juang**, (1986), An introduction to Hidden Markov Models, *IEEE ASSP Magazine*, vol. 3, pp. 4 – 16.

[27] **T. F. Quatieri**, (1979), Minimum and mixed phase speech analysis synthesis by adaptive homomorphic deconvolution, *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 27, pp. 328–335.

[28] **C. G. M. Fant**, *Speech Sounds and Features.* MIT Press: Cambridge,MA., 1973.

[29] **R. W. B. Stephens and A. E. Bate**, *Acoustics and Vibrational Physics.* New York: St. Martins Press, 1966.

[30] **J. W. Pitton, L. E.Atlas, and P. J. Loughlin**, (1994), Applications of positive time frequency distributions to speech processing, *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 2(4), pp. 554–566.

[31] **A. P. Reilly and B. Boashash**, (1992), A comparison of time frequency signal analysis techniques with application to speech recognition, *Proceedings of the SPIE The International Society for Optical Engineering*, vol. 1770, pp. 339–350.

[32] **D. Rainton and S. J. Young**, (1992), Time frequency spectral estimation of speech, *Computer Speech and Language*, vol. 6(1), pp. 15–36.

[33] **D. B. Paul and B. F. Necioglu**, (1993), The Lincoln large vocabulary stack decoder HMM CSR, *Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing*, vol. II(2), pp. 660–663.

[34] **H. A. Murthy, F. Beaufays, and L. P. Heck**, (1999), Robust Text-Independent Speaker Identification over Telephone Channels, *IEEE Trans. Signal Processing*, vol. 7(5), pp. 554–568.

[35] **J. Makhoul**, (1975), Linear prediction: A tutorial review, *Proceedings of IEEE*, vol. 63, pp. 561–580.

[36] **B. Yegnanarayana**, (1978), Formant extraction from linear prediction phase spectrum, *J. Acoust. Soc. Amer.*, pp. 1638–1640.

[37] **H. A. Murthy and V. R . R. Gadde**, (2003), The Modified group delay function and its application to phoneme recognition, *Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing*, vol. I, Hong Kong, April, pp. 68–71.

[38] **B. Yegnanarayana, D. K. Saikia, and T. R. Krishnan**, (1984), Significance of group delay functions in signal reconstruction from spectral magnitude or phase, *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 32, no. 3, pp. 610–622.

[39] **K. V. M. Murthy and B. Yegnanarayana**, (1989), Effectiveness of representation of signals through group delay functions, *Elsevier Signal Processing*, vol. 17, pp. 141–150.

[40] **L. R. Rabiner and R. W. Schafer**, *Digital Processing of Speech Signals.* Englewood Cliffs, NJ: Prentice-Hall, 1978.

[41] **P. Yip and K. R. Rao**, *Discrete Cosine Transform: Algorithms, Advantages and Applicatons.* USA: Academic Press, 1997.

[42] **P. Alexandre and P. Lockwood**, (1993), Root Cepstral Analysis: a Unified view. Application to Speech Processing in Car Noise Environments, *Speech Communication*, vol. 12(3), pp. 277–288.

[43] **R. Sarikaya and J. H. L. Hansen**, (2001), Analysis of the root-cepstrum for acoustic modeling and fast decoding in speech recognition, *Eurospeech*, Denmark, September, pp. 687–690.

[44] **H. Hermansky and N. Morgan**, (1994), RASTA processing of speech, *IEEE Trans. Acoust., Speech, Signal Processing*, pp. 578–589.

[45] **K. Fukunaga**, *Introduction to Statistical Pattern Recognition.* Boston: Academic Press, 1990.

[46] **R. M. Hegde, H. A. Murthy and V. R. R. Gadde**, (2004), Application of the modified group delay function to speaker identification and discrimination, *Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing*, vol. 1, Montreal, May, pp. 517–520.

[47] **Y. K. Muthusamy, R. A. Cole, and B. T. Oshika**, (1992), The OGI multilanguage telephone speech corpus, *Proceedings of Int. Conf. Spoken Language Processing*, Oct, pp. 895–898.

[48] **J. W. Sammon, Jr.**, (1969.), A nonlinear mapping for data structure analysis, *IEEE Trans. Comput.*, vol. C-18(5), pp. 401–409.

[49] **A. Acero**, *Acoustical and Environmental Robustness in Automatic Speech Recognition.* PhD dissertation, CMU, Pittsburgh, PA, U.S.A., 1990.

[50] **H. A. Murthy and B. Yegnanarayana**, (1989), Speech processing using group delay functions, *Elsevier Signal Processing*, vol. 17, pp. 141–150.

[51] **V. R. R. Gadde, A. Stolcke, J. Z. D. Vergyri, K. Sonmez, and A. Venkatraman**, *The SRI SPINE 2001 evaluation system.* Menlo Park, CA.: SRI, 2001.

[52] **T. Nagarajan, V. K. Prasad, and H. A. Murthy**, The minimum phase signal derived from the root cepstrum, *IEE Electronic Letters*, vol. 39, pp. 941–942.

[53] **A. Janin, D. Ellis, and N. Morgan**, (1999), Multi-stream speech recognition: Ready for prime time ?, *Proceedings of EUROSPEECH*, September, pp. 591–594.

[54] **S. Okawa, E. Bocchieri, and A. Potamianos**, (1998), Multiband speech recognition in noisy environments, *Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing*, vol. 2, May, pp. 641–644.

[55] **D. Ellis**, (2000), Stream combination before and/or after the acoustic model, *Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing*, vol. 3, pp. 1635–1638.

[56] **B. Kingsbury and N. Morgan**, (1997), Recognizing reverberant speech with RASTA-PLP, *Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing*, vol. 2, pp. 1259–1262.

[57] **S. Wu, B. Kingsbury, N. Morgan, and S. Greenberg**, (1998), Incorporating information from syllable-length time scales into automatic speech recognition, *Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing*, vol. 2, May, pp. 721–724.

[58] **K. Kirchoff and J. Bilmes**, (1999), Dynamic classifier combination in hybrid speech recognition systems using utterance level confidence values, *Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing*, April.

[59] **A. Arai and S. Greenberg**, (1998), Speech intelligibility in the presence of cross-channel spectral asynchrony, *Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing*, vol. 2, pp. 929–932.

[60] **O. Ghitza**, (1994), Auditory models and human performance in tasks related to speech coding and speech recognition, *Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing*, vol. 2(1), January, pp. 115–132.

[61] **X. Yang, K. Wang, and S. A. Sharma**, (1992), Auditory representations of acoustic signals, *IEEE Trans. Inform. Theory*, vol. 38(2), pp. 824–839.

[62] **S. Greenberg**, (1999), Speaking in shorthand - a syllable centric perspective for understanding pronunciation variation, *ESCA Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition*, pp. 47–56.

[63] **C. Antoniou and T. Reynolds**, (1999), Using modular/ensemble neural networks for the acoustic modeling of speech, *IEEE Workshop on ASRU*, Keystone, Colorado.

[64] **C. Antoniou and T. Reynolds**, (2000), Acoustic modeling using modular/ensemble combinations of heterogeneous neural networks, *Proceedings of Int. Conf. Spoken Language Processing*, pp. 282–285.

[65] **A. Sharma, D. Ellis, S. Kajarekar, P. Jain, and H. Hermansky**, (2000), Feature extraction using non-linear transformation for robust speech recognition on the aurora database, *Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing*, vol. 2, pp. 1117–1120.

[66] **S. Greenberg and B. E. D. Kingsbury**, (1997), The modulation spectrogram : In pursuit of an invariant representation of speech, *Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing*, pp. 1647–1650.

[67] **H. Hermansky and S. Sharma**, (1999), Temporal patterns (TRAPS) in ASR of noisy speech, *Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing*, pp. 289–292.

[68] **G. Cook, J. Christie, D. Ellis, E. Fosler-Lussier, Y. Gotoh, B. Kingsbury, N. Morgan, S. Renals, A. Robinson, and G. Williams**, (1999), The sprach system for the transcription of broadcast news, *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop.*

[69] **M. L. Shire**, (2001), Multi-stream ASR trained with heterogeneous reverberant environments, *Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing*, vol. 1, pp. 253–256.

[70] **K. Kirchoff**, (1998), Combining articulatory and acoustic information for speech recognition in noisy and reverberation environments, *Proceedings of Int. Conf. Spoken Language Processing*, pp. 891–894.

[71] **K. Kirchoff, and J. Bilmes**, (2000), Combination and joint training of acoustic classifiers for speech recognition, *ISCA ITRW Workshop on Automatic Speech Recognition (ASRU 2000)*, pp. 17–23.

[72] **H. Christensen, B. Lindberg, and O. Andersen**, (2000), Employing heterogeneous information in a multi-stream framework, *Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing*, vol. III, pp. 1571–1574.

[73] **H. Meinedo and J. Neto**, (2000), Combination of acoustic models in continuous speech recognition hybrid systems, *Proceedings of Int. Conf. Spoken Language Processing*, vol. 2, pp. 931–934.

[74] **M. Heckmann, F. Berthommier, and K. Kroschel**, (2001), Optimal weighting of posteriors for audio-visual speech recognition, *Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing*, vol. 1, pp. 161–164.

[75] **N. Mirghafori**, *A Multi-Band Approach to Automatic Speech Recognition.* PhD dissertation, ICSI, Berkeley, California, U.S.A., 1999.

[76] **A. Cerisara**, (2000), Asynchrony in multi-band speech recognition, *Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing*, vol. 2, pp. 1121–1124.

[77] **S. Dupont**, *Etudes et Developpement de Nouveaux Paradigmes pour la Reconnaissance Robuste de la Parole.* PhD dissertation, Laboratoire TCTS, Universite de Mons, Belgium, 2000.

[78] **H. Hermansky, S. Tibrewala, and M. Pavel**, (1996), Towards ASR on partially corrupted speech, *Proceedings of Int. Conf. Spoken Language Processing*, pp. 462–465.

[79] **J. Allen**, (1994), How do humans process and recognize speech ?, *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 2(4), pp. 567–577.

[80] **H. Bourlard, S. Bengio, and K. Weber**, *Towards Robust and Adaptive Speech Recognition Models.* Switzerland: IDIAP Research Report 02-01, 2002.

[81] **L. R. Rabiner and B. H. Juang**, *Fundamentals of Speech Recognition.* New Jersey: Prentice Hall, 1993.

[82] **X. D. Huang, Y. Ariki, and M. A. Jack**, *Hidden Markov Models for Speech Recognition.* Edinburgh: Edinburgh Univ. Press, 1990.

[83] **B. H. Juang**, (1985), Maximum-likelihood estimation for mixture multivariate stochastic observations of Markov chains, *AT & T Bell Laboratories Technical Journal*, vol. 64, no. 6, pp. 1235–1249.

[84] **J. R. Deller, J. G. Proakis, and J. H. L. Hansen**, *Discrete-Time Processing of Speech Signals.* New York: MacMillan series for Prentice-Hall,, 1993.

[85] **L. E. Baum, T. Petrie, G. Soules, and N. Weiss**, (1970), A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains, *Ann. Math. Statist.*, vol. 41, no. 1, pp. 164–171.

[86] **A. J. Viterbi**, (1967), Error bounds for convolutional codes and an asymptotically optimal decoding algorithm, *IEEE Trans. Inform. Theory*, vol. 13, no. 2, pp. 260–269.

[87] **G. D. Forney**, (1973), The Viterbi algorithm, *Proc. IEEE*, vol. 61, pp. 268–277.

LIST OF PAPERS SUBMITTED ON THE BASIS OF THIS THESIS

REFEREED JOURNALS

1. **Rajesh M. Hegde, Hema A. Murthy, and Venkata Ramana Rao Gadde**, "Significance of The Modified Group Delay Feature in Speech Recognition", *IEEE Transactions on Speech and Audio Processing*, (Revised and Re-submitted).
2. **Rajesh M. Hegde, Hema A. Murthy, and Venkata Ramana Rao Gadde**, "Significance of Joint Features Derived from The Modified Group Delay Function in Speech Processing", *EURASIP Journal on Applied Signal Processing*, (Revised and Re-submitted).
3. **Rajesh M. Hegde and Hema A. Murthy**, "Significance of Multi-Stream and Selective-Band Feature Combination in ASR", *Speech Communication*, (Revised and Re-submitted).

REFEREED INTERNATIONAL CONFERENCES

1. **Rajesh M. Hegde, Hema A. Murthy, and Venkata Ramana Rao Gadde**, "Speech Processing using Joint Features Derived from The Modified Group Delay Function," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP-2005*, Philadelphia, U.S.A, Vol 1, pp. 541-544, Mar. 2005.
2. **Rajesh M. Hegde and Hema A. Murthy**, "Automatic Language Identification and Discrimination using The Modified Group Delay Feature", in *Proceedings of International Conference on Intelligent Sensing and Information Processing, ICISIP-2005*, Chennai, pp. 395 - 399, Jan. 2005.
3. **Rajesh M. Hegde and Hema A. Murthy**, "An Alternative Representation of Speech using The Modified Group Delay Feature", in *Proceedings of International Conference on Signal Processing and Communications, SPCOM-2004*, Bangalore, Su C1.1, Dec. 2004.

4. **Rajesh M. Hegde and Hema A. Murthy**, "Cluster and Intrinsic Dimensionality Analysis of The Modified Group Delay Feature for Speaker Classification", *Lecture Notes in Computer Science,* **LNCS 3316***, pp. 1172 - 1178, Springer Verlag, From Proceedings of the ICONIP-2004*, Calcutta, India, Nov. 2004.
5. **Rajesh M. Hegde, Hema A. Murthy, and Venkata Ramana Rao Gadde**, "Continuous Speech Recognition using Joint Features derived from The Modified Group Delay Function and MFCC," in *Proceedings of the INTERSPEECH 2004 - ICSLP*, Jeju, S. Korea, Vol. 2, pp. 905-908, Oct. 2004.
6. **Rajesh M. Hegde, Hema A. Murthy, and Venkata Ramana Rao Gadde**, "The Modified Group Delay Feature : A New Spectral Representation of Speech," in *Proceedings of the INTERSPEECH 2004 - ICSLP*, Jeju, S. Korea, Vol. 2, pp. 913-916, Oct. 2004.
7. **Rajesh M. Hegde, Hema A. Murthy, and Venkata Ramana Rao Gadde**, "Application of the Modified Group Delay Function to Speaker Identification and Discrimination," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP-2004*, Montreal, Canada, Vol 1, pp. 517-520, May 2004.
8. **Rajesh M. Hegde and Hema A. Murthy**, "Speaker Identification using The Modified Group Delay Feature," in *Proceedings of International Conference on Natural Language Processing, ICON 2003*, Mysore, India, pp. 159-167, Dec. 2003.
9. **T. Nagarajan, Hema A. Murthy, and Rajesh M. Hegde**, "Segmentation of Speech into Syllable-like Units", *Proc. EUROSPEECH-2003*, Geneva, Switzerland, Sep. 2003, pp.2893-2896.
10. **T. Nagarajan, Hema A. Murthy, and Rajesh M. Hegde**, "Group Delay based Segmentation of Spontaneous Speech into Syllable-like Units", *ISCA and IEEE workshop on Spontaneous Speech Processing and Recognition, SSPR-2003*, Tokyo, Japan, Apr. 2003, pp.115-118.

CURRICULUM VITAE

1 **Name**: Rajesh Mahanand Hegde

2 **Permanent Address**:

S/O M. M. Hegde
Plot no. 2, Shakti Colony
Visvesvaranagar
Hubli - 580032
Karnataka, India

3 **Educational Qualification**:

PhD (2005)	Indian Institute of Technology, Madras
Computer Science and Engg.	Chennai - 600 036
	India
M.E. (1998)	Bangalore University
Electronics Engg.	Karnataka, India
B.E. (1992)	Mysore University
Instrumentation Engg.	Karnataka, India

DOCTORAL COMMITTEE MEMBERS

1 **Chairperson**: **Prof. S. Raman**

2 **Guide**: **Dr. Hema A. Murthy**

3 **Members** :

Prof. Timothy A. Gonsalves
Dept. of Computer Science & Engg.

Prof. C. Siva Ram Murthy
Dept. of Computer Science & Engg.

Prof. Bhaskar Ramamurthy
Dept. of Electrical Engg.

Dr. K. Giridhar
Dept. of Electrical Engg.