FEATURE ANALYSIS AND COMPENSATION FOR SPEAKER RECOGNITION UNDER STRESSED CONDITION

A Thesis Submitted in

Partial Fulfillment of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

By

G. SENTHIL RAJA

Lakshminath Bezbaroa Central Library
IIT Guwahati

DEPARTMENT OF ELECTRONICS AND COMMUNICATION ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY GUWAHATI

JULY, 2007

Indian Institute of Technology Guwahati

This work is dedicated to

my guide

Dr. S. DANDAPAT

for his guidance

my parents

G. GUNASEELA BOOPATHY

and

G. CHANDRA

for their encouragement

my brother

G. SARAVANA BOOBALAN

for his support

my wife

V. TAMILARASI

for her help during the preparation of thesis

TH-1853_02610201

Certificate

This is to certify that the thesis entitled "**FEATURE ANALYSIS AND COMPENSATION FOR SPEAKER RECOGNITION UNDER STRESSED CONDITION**", submitted by G. Senthil Raja, a research student in the *Department of Electronics and Communication Engineering , Indian Institute of Technology, Guwahati*, for the award of the degree of **Doctor of Philosophy**, is a record of an original research work carried out by him under my supervision and guidance. The thesis has fulfilled all requirements as per the regulations of the Institute and in my opinion has reached the standard needed for submission. The results embodied in this thesis have not been submitted to any other University or Institute for the award of any degree or diploma.

Dated: 16.07.2007
Guwahati.

Dandapat

Dr. Samarendra Dandapat
Professor
Dept. of Electronics and Communication Engg.
Indian Institute of Technology Guwahati
Guwahati - 781 039
India.

TH-1853_02610201

Indian Institute of Technology Guwahati

Acknowledgements

I feel it is a great privilege to express my deepest and most sincere gratitude and respect to my supervisor, Dr. Samarendra Dandapat for his suggestions, constant encouragement and support during the course of the thesis work. I am also grateful to the other members of my doctoral committee, namely Dr. J. S. Sahambi (Chairman), Dr. S. R. M. Prasanna, and Dr. P. Viswanath for their valuable comments on my work. I thank Prof. P. K. Bora and Dr. K. Sreenivasa Rao for their comments on my thesis work. I would also like to thank the other faculty members of the department, namely Prof. A. Mahanta, Prof. A. K. Gogoi, Dr. S. Majhi, Dr. C. Mahanta, Dr. H. B. Nemade, Dr. R. Bhattacharjee, Dr. Roy P. Paily, Dr. A. Mitra, Dr. K. Rakesh Singh and Dr. P. R. Sahu for their encouragement and help.

I am also grateful to all the members of the Research and technical staff of the department, namely L. N. Sharma, Sanjib Das, Usha Bharali, S. Josephine, Utpal Kumar Sharma, Pranab Jyoti Goswami, Sidananda Sonowal and Jharna Rani Rabha without whose help I could not have completed this assignment. I thank all my fellow research students and M. Tech students for their cooperation. My thanks also go out to all my friends, namely P. Vinod, P. Suresh kumar, J. W. J. Bosco, M. S. Manikandan, M. Karthikeyan, P. Rajesh Kanna, M. Sivasankar, S. Manikandan, Janmejaya rout, and N. Suresh kumar who have made my stay at Guwahati a memorable period of my life.

I would like to thank my family for the support they provided me through my entire life and in particular, I must acknowledge my grand father K. Venugopal, aunty G. Raja Kumari uncle R. Govindaraj, father-in-law M. Venkataraman and mother-in-law C. Kalaiselvi without those love, encouragement and assistance, I would not have finished. Finally, I must express my heart-felt gratitude to my wife V. Tamilarasi for her encouragement and support in final stage of my thesis.

(G. Senthil Raja)

TH-1853_02610201

Indian Institute of Technology Guwahati

Abstract

This thesis documents our investigations on feature analysis and design of compensators for speaker recognition under stressed conditions. Any condition that causes a speaker to vary his or her speech production from normal or neutral condition is called stressed speech condition. Stressed speech is induced by emotion, high workload, sleep deprivation, frustration and environmental noise. In stressed condition, the characteristics of speech signal are different from that of normal or neutral condition. Due to changes in speech signal characteristics, the performance of speaker recognition system may degrade under stressed speech conditions. In this work, the problem of speaker recognition under stressed condition has been dealt with three broad approaches.

First, six speech features (mel-frequency cepstral coefficients (MFCC), linear prediction (LP) coefficients, linear prediction cepstral coefficients (LPCC), reflection coefficients (RC), arc-sin reflection coefficients (ARC) and log-area ratios (LAR)), which are widely used for speaker recognition, are analyzed for evaluation of their characteristics under stressed condition. Statistical techniques such as Probability density function (pdf) characteristics, F-ratio test and Kolmogorov-Smirnov (K-S) test are used for evaluation of speaker discrimination capability of the features under stressed condition. Two classifiers, Vector Quantization (VQ) classifier and Gaussian Mixture model (GMM), are used to evaluate the speaker recognition results with different speech features. Relative ranking of the features based on their performance in speaker recognition under stressed condition is evaluated. This analysis help select the best feature set for speaker recognition under stressed condition.

Secondly, sinusoidal model is used for analysis of speaker recognition system under stressed condition. Sinusoidal model has three different parameters or features. These are amplitude, frequency and phase features. With the assumption that these three features may carry different information of a speaker, the performance of the system can be improved by suitably combining the three features at the classifier stage. In this work, we propose evaluation of confusion probability from confusion matrix for analysis of independent information with sinusoidal amplitude feature (SAF) and sinusoidal frequency feature (SFF). Results show that there may be considerable amount of independent information between the sinusoidal amplitude feature (SAF) and sinusoidal frequency feature (SFF). Improvement in speaker verification is observed by combining the amplitude and the frequency features.

Finally, four novel compensation techniques are proposed and evaluated for improvement of speaker recognition under stressed condition. The compensation techniques are speaker and stressed informa-

TH-1853_02610201

tion based compensation (SSIC), compensation by removal of stressed vectors (CRSV), cepstral mean normalization (CMN) and combination of MFCC and sinusoidal amplitude (CMSA) features. Compensation techniques improve speaker recognition results for speech signals under stressed conditions. Speech data from SUSAS database corresponding to four different emotions or stressed conditions, Angry, Lombard, Question and Neutral, are used for analysis of speaker recognition under stressed condition.

TH-1853_02610201

Indian Institute of Technology Guwahati

Contents

List of Figures iii

List of Tables v

Abbreviations vii

Mathematical Notations ix

1 Introduction 1
- 1.1 Overview of Speaker Recognition . . . 2
  - 1.1.1 Text-dependent speaker recognition . . . 4
  - 1.1.2 Text-independent speaker recognition . . . 4
- 1.2 Major Tasks of Speaker Recognition Systems . . . 5
  - 1.2.1 Feature extraction . . . 6
  - 1.2.2 Classifiers . . . 8
  - 1.2.3 Similarity matching . . . 10
  - 1.2.4 Decision logic . . . 13
- 1.3 Speaker Recognition and Stressed Speech . . . 14
  - 1.3.1 Analysis and recognition of stressed speech . . . 15
- 1.4 Scope of the Present Work . . . 17
- 1.5 Organization of the thesis . . . 19

2 Review of Speech Features and Speaker Recognition 21
- 2.1 Speech Features for Speaker Recognition . . . 22
  - 2.1.1 Linear prediction model based features . . . 23
  - 2.1.2 Mel-Frequency cepstral coefficients (MFCC) . . . 25
- 2.2 Statistical Feature Evaluation . . . 26
  - 2.2.1 Probability density function characteristics . . . 27
  - 2.2.2 Analysis of variance . . . 27
  - 2.2.3 Kolmogorov-Smirnov (K-S) test . . . 30
- 2.3 Classifiers for Speaker Recognition . . . 33
  - 2.3.1 Vector quantization (VQ) . . . 33
  - 2.3.2 Gaussian mixture model . . . 33
  - 2.3.3 Speaker recognition system . . . 35
- 2.4 Motivation for the Present Work . . . 38

TH-1853_02610201

3 Evaluation of Speech Features for Speaker Recognition under Stressed Condition **44**
3.1 Statistical Evaluation of Speech Features 45
3.1.1 Probability density function characteristics 46
3.1.2 Analysis of variance 49
3.1.3 Kolmogorov-Smirnov Test 53
3.2 Feature Analysis using Classifiers 56
3.2.1 Vector quantization (VQ) classifier 56
3.2.2 Gaussian Mixture Model (GMM) based classifier 62
3.3 Relative Ranking of the Features 65

4 Speaker Information Analysis using Sinusoidal Model Features **72**
4.1 Sinusoidal Modeling of Speech Signal 73
4.1.1 Speaker discrimination with sinusoidal features 76
4.2 Statistical Analysis of Sinusoidal Model Features 78
4.3 Speaker Recognition with Sinusoidal Model 80
4.3.1 Independent information analysis 83
4.3.2 Speaker recognition under stressed condition 86
4.4 Combination of MFCC and Sinusoidal Features 90

5 Stress Compensation for Speaker Recognition **93**
5.1 Speaker and Stress Information based Compensation(SSIC) 94
5.2 Compensation by Removal of Stressed Vectors (CRSV) 101
5.3 Combination of MFCC and Sinusoidal Amplitude (CMSA) Features 103
5.4 Cepstral Mean Normalization (CMN) 105
5.5 Selection of Compensation by Stress Recognition (SCSR) 107
5.5.1 Stress recognition 108

6 Conclusion **111**
6.1 Scope of future work 116

References **118**

TH-1853_02610201

List of Figures

1.1 Block diagram of speaker recognition system 4
1.2 Speaker recognition systems, (a) speaker identification system and (b) speaker verification system. 5

2.1 Mel-frequency filter bank 26
2.2 Probability density function characteristics 28
2.3 F-ratio values of six speech features evaluated for nine speakers 30
2.4 Maximum KS-test distance 32
2.5 Summation of Mean Square Error between training utterance and training Models(Codebook) 36
2.6 Speech signals (word: 'break') recorded from one speaker and their Spectrograms in two different emotions (a) Neutral (b) Angry 39
2.7 Histograms of MFCC coefficients from Speech signals (word: 'break') recorded from one speaker in two different emotions (a) Neutral (b) Angry 40

3.1 Spectrograms of speech signals (word: 'break') recorded from one speaker in two different emotions (a) Neutral (Non Pre-emphasized) (b) Neutral (Pre-emphasized) (c) Angry (Non Pre-emphasized) (d) Angry (Pre-emphasized) 46
3.2 Probability density function of 12 LP cepstral coefficients (LPCC) for four speakers . . 47
3.3 Probability density function of 12 LPCCs of four emotions data of a single speaker . . 48
3.4 F-ratio values of six features of Neutral speech (Solid line - non pre-emphasized speech, Dashed line - pre-emphasized speech) 51
3.5 F-ratio values of six features of stressed speech (Solid line - non pre-emphasized speech, Dashed line - pre-emphasized speech) 52
3.6 F-ratio values of six features (Solid line - speaker, Dashed line - stress) 53
3.7 Genuine (diamond symbol) and Imposter (circular symbol) score of Log area ratios . . 60
3.8 Genuine (diamond symbol) and Imposter (circular symbol) score of Mel-frequency cepstral coefficients 60
3.9 Genuine (diamond symbol) and Imposter (circular symbol) score of Cepstral coefficients 61
3.10 Genuine (diamond symbol) and Imposter (circular symbol) score of Arc-sin reflection coefficients 61
3.11 Genuine (diamond symbol) and Imposter (circular symbol) score of Reflection coefficients 62
3.12 Genuine (diamond symbol) and Imposter (circular symbol) score of Linear prediction coefficients 62
3.13 DET curve for different features under neutral condition 63
3.14 DET curve for different features under stressed condition 64

TH-1853_02610201

4.1 Sinusoidal model feature extraction 74
4.2 (a) Voiced speech signal, (b) Reconstructed voiced speech signal 75
4.3 (a) Unvoiced speech signal, (b) Reconstructed unvoiced speech signal 75
4.4 Mean square error versus order for a voiced speech signal 76
4.5 Amplitude contour 77
4.6 Frequency contour 77
4.7 Probability density function of amplitude feature for four speakers 79
4.8 Probability density function of frequency feature for four speakers 79
4.9 Probability density function of phase features for four speakers 80
4.10 F-ratios of sinusoidal features. Solid line shows F-ratios when evaluated between the speakers and dashed line shows F-ratios when evaluated between the four emotions. . . 81
4.11 Significant peak selection from magnitude spectrum 82
4.12 Speaker Identification rate versus order 83
4.13 Equal error rate of combined confidence score of SAF and SFF 86
4.14 Equal error rate of combined confidence score for different combination. 91
4.15 Equal error rate of MFCC and SAF for stressed speech. 92

5.1 Speaker recognition system with a stress compensation technique 94
5.2 F-ratio values of MFCC feature 96
5.3 Multiplication factors for the error distance in the codebook 98
5.4 Multiplication factors of error distances for corresponding coefficients 99
5.5 Average speaker identification with different values of stress weights 100
5.6 Multiplication factors of error distances 100
5.7 Compensation by removal of stressed vector 102
5.8 Stress compensation using combination of MFCC and SAF features 104
5.9 Speaker identification results of four emotions 104
5.10 Average speaker identification results 105
5.11 Block diagram of selection of compensation by Stress recognition (SCSR) system . . 107

TH-1853_02610201

Indian Institute of Technology Guwahati

List of Tables

2.1 Speaker identification results of YOHO database (106 male speakers, 256 size codebook) for different speech features . . . 37
2.2 Speaker identification results of SUSAS database (9 speakers, 32 size codebook) for different speech features . . . 37
2.3 Equal error rate (%) and false rejection rates at false acceptance rates of 0.1% and 0.01% for YOHO database . . . 37
2.4 Speaker identification rate (%) for different emotions . . . 41

3.1 Average F-ratio values of different features . . . 50
3.2 Divergence values of six features . . . 54
3.3 Average of maximum distances between the two possible speakers CDFs . . . 55
3.4 Average of Maximum distances between CDFs of a speaker's reference stress condition and all other stress conditions . . . 56
3.5 Speaker identification rate using VQ classifier with Neutral emotion codebook (pre-emphasized speech) . . . 58
3.6 Speaker identification rate using VQ classifier with Neutral emotion codebook (non pre-emphasized) . . . 59
3.7 Speaker recognition results under non pre-emphasized (Neutral emotion) condition . . . 63
3.8 Speaker recognition results under stressed condition with non pre-emphasized (Neutral emotion) speaker models . . . 64
3.9 Speaker identification rate (%) for different emotions using MFCC . . . 65
3.10 Speaker identification rate (%) for non pre-emphasized speech . . . 68
3.11 Speaker identification rate (%) for pre-emphasized speech . . . 69
3.12 Speaker identification rate using GMM classifier for pre-emphasized speech (Training: Neutral) . . . 70
3.13 Feature analysis for speaker recognition under stressed condition (pre-emphasized) . . . 70
3.14 Position probability of six speech features for speaker recognition under stressed condition . . . 71
3.15 Position probability of six speech features for speaker recognition under stressed condition . . . 71

4.1 Average F-ratio, average KS-test distance and Divergence values of sinusoidal model features under stressed condition . . . 81
4.2 Speaker recognition results under stressed condition with normal(Neutral emotion) speaker models . . . 87
4.3 Speaker identification rate (%) for sinusoidal amplitude feature with log magnitude squared coefficients. . . . 87

TH-1853_02610201

4.4 Speaker identification rate (%) for sinusoidal amplitude feature with amplitude values coefficients. 88
4.5 Speaker identification rate (%) for sinusoidal frequency features. 88
4.6 Feature analysis for speaker recognition under stressed condition (Pre-emphasized) . . 89

5.1 F-ratio of fourteen MFCC coefficients(%) 96
5.2 Speaker identification rate (%) with F-ratio test analysis 97
5.3 Speaker identification rate (%) for different multiplication factors with the error distances in the codebook (14 coefficients) 99
5.4 Speaker identification rate (%) for different multiplication factors with the error distances in the codebook (12 coefficients) 101
5.5 Speaker identification rate (%) 102
5.6 Maximum speaker identification results (%) for combined MFCC and SAF features . . 103
5.7 Speaker identification rate (%) 106
5.8 SI results (%) of different compensators for SUSAS database 108
5.9 Stress classification using VQ classifier 109
5.10 SI results (%) of different compensators for SUSAS database 110

TH-1853_02610201

Abbreviations

ARC	Arc-sin Reflection Coefficients
$ARMA$	Autoregressive Moving Average
CC	Combined confidence score
CDF	Cumulative Distribution Function
CMN	Cepstral Mean Normalization
$CMSA$	Combination of MFCC and Sinusoidal Amplitude feature
CNC	Common Neutral Code book
$CRSV$	Compensation by Removal of Stressed Vector
$DHMM$	Discrete Hidden Morkov Model
EER	Equal Error Rate
EM	Expectation Maximization
FAR	False Acceptance Rate
FRR	False Rejection Rate
$F-ratio$	Fisher's discriminant ratio
FFT	Fast Fourier Transform
GMM	Gaussian Mixture Model
HMM	Hidden Markov Model
KS	Kolmogrov-Smirnov test
LP	Linear Prediction
$LPCC$	Linear Prediction Cepstral Coefficients
LAR	Log Area Ratios
LSP	Line Spectrum Pair
$MFCC$	Mel-Frequency Cepstral Coefficients
PDF	Probability Density Function
PLP	Perceptual Linear Prediction
RC	Reflection Coefficients
ROC	Receiver Operating Characteristics

TH-1853_02610201

SE	Stress Effect
SMF	Sinusoidal Model Features
SAF	Sinusoidal Amplitude Features
SFF	Sinusoidal Frequency Features
SPF	Sinusoidal Phase Features
$SUSAS$	Speech Under Simulated and Actual Stress
$STFT$	Short Time Fourier Transform
$SMSE$	Sum of Mean Square Error
$SCSR$	Selection of Compensation by Stress Recognition
SI	Speaker Identification
$SSIC$	Speaker and Stressed Information based Compensation
SV	Speaker verification
VQ	Vector Quantization

TH-1853_02610201

Mathematical Notations

a_h	h^{th} prediction coefficient
$A_j^{(k)}, f_j^{(k)}, \phi_j^{(k)}$	Amplitude, frequency and phase values of j^{th} sinusoid of k^{th} frame
B	interspeaker covariance matrix
c_m	m^{th} Cepstral coefficient
$\vec{c}$	Codebook vector
C	Codebook
C_{fe}	Confusion matrix of fe^{th} feature
CS_{fe}	Confidence score of fe^{th} feature
CS	Combined Confidence score
d	Quadratic distance
Div	Divergence
f	Actual frequency
f_i	Represents F-ratio value of the i^{th} coefficient
F_s	Sampling frequency
F	Fisher's discriminant ratio
$f(x)$	Probablity distribution function
$F(x)$	Cumulative distributive function
g_m	Log Area Ratio
$g_m^{'}$	Arc-sin reflection coefficient
i	Speaker idex
I_{IN}	Higher independent information
I	Total number of speakers
k	Frame index
k_m	Reflection coefficient
L	Number of sinusoids

TH-1853_02610201

M	Number of frames
$mel(f)$	Mel-frequency
N	Length of frame
mix	Gaussian mixture index
n	Speech sample index
p	Order of linear prediction
P_{fe}	Probability Confusion matrix of fe^{th} feature
P_D	Measure of independent information
$p_{mix}(x)$	unimodel gaussian density
$\tilde{s}[n]$	Speech frame
T	Total number of vectors
U_{ij}	i^{th} speaker j^{th} training utterance
w	intraspeaker covariance matrix
W	Positive definite matrix
w_{mix}	Mixture weights
$\hat{x}[n]$	n^{th} predicted speech sample
$\vec{x}$	Test vector
X	Test vectors
$x_{\beta}^{(i)}$	$x_{\beta}^{(i)}$ is the feature value from the β^{th}utterance spoken by the i^{th} speaker
x, y	N-dimensional templates
α	Weighting factor
σ^2	Gain term
β	Utterance index
$\bar{\mu}$	Estimated overall mean value
μ_i	i^{th} Mean value
Σ_{mix}	covariance matrix
λ	density model or speaker model

TH-1853_02610201

Indian Institute of Technology Guwahati

Chapter 1

Introduction

This thesis documents our investigations on feature analysis and design of compensators for speaker recognition under stressed speech conditions. Any condition that causes a speaker to vary his or her speech production from normal or neutral condition is called stressed speech condition. Stressed speech is induced by emotion, high workload, sleep deprivation, frustration and environmental noise. In stressed condition, the characteristics of speech signal are different from that of normal or neutral condition. Due to changes in speech signal characteristics, performance of the speaker recognition system may degrade under stressed speech conditions. In this work, the problem of speaker recognition under stressed condition has been dealt with three broad approaches. First, six speech features (mel-frequency cepstral coefficients (MFCC), linear prediction (LP) coefficients, linear prediction cepstral coefficients (LPCC), reflection coefficients (RC), arc-sin reflection coefficients (ARC) and log-area ratios (LAR)), which are widely used for speaker recognition, are analyzed for evaluation of their characteristics under stressed condition. Statistical techniques such as Probability density function (pdf) characteristics, F-ratio test and Kolmogorov-Smirnov (K-S) test are used for evaluation of speaker discrimination capability of the features under stressed condition. Two classifiers, Vector Quantization (VQ) classifier and Gaussian Mixture model (GMM), are used to evaluate the speaker recognition results with different speech features. Relative ranking of the features based on their performance in speaker recognition under stressed condition is evaluated. This analysis help select the best feature set for speaker recognition under stressed condition. Secondly, sinusoidal model is used for analysis of speaker recognition system under stressed condition. Sinusoidal model has three different parameters or features. These are amplitude, frequency and phase features. With the assumption that these three features may carry different information of a speaker, the performance of the system can be improved by suitably combining the three features at the classifier stage. Finally, four novel compensation techniques are proposed and eval-

TH-1853_02610201

uated for improvement of speaker recognition under stressed condition. The compensation techniques are speaker and stressed information based compensation (SSIC), compensation by removal of stressed vectors (CRSV), cepstral mean normalization (CMN) and combination of MFCC and sinusoidal amplitude (CMSA) features. Speech data from SUSAS database corresponding to four different emotions or stressed conditions, Angry, Lombard, Question and Neutral, are used for analysis of speaker recognition under stressed condition. In the following sections, some major results and related discussions in different chapters of the thesis are briefly presented.

1.1 Overview of Speaker Recognition

Recognition of speaker from his or her voice (speech signal) by machine is known as speaker recognition. Speaker recognition is a well known area of human-computer interaction. Speaker recognition has a variety of applications. Some of the applications in the access control are, voice dialing, banking transactions through telephone network, teleshopping, database access service, reservation systems, voice mail, security control for confidential information access, and remote access to computers. Another major application is in forensic sciences where voice information is used for identification of criminals.

Speech signal carries linguistic, speaker and environmental information. Speech signal is a slowly time varying signal. Over a sufficiently short period of time (between 5 and 30 msec), its characteristics are fairly stationary . However over long periods of time, the signal characteristics change to reflect different speech sounds being spoken. Speech is classified as voiced and unvoiced depending on the nature of excitation [1]- [2]. In case of voiced speech, the excitation of the vocal tract originates at the glottis and it is due to the periodic vibration of the vocal cords. The frequency of vibration, or pitch, is controlled by the tension in the vocal cords and the air pressure from the lungs. Typical pitch values lie in the range 50-500 Hz for adults, and can rise to 1000 Hz in children. Due to its periodic nature, the spectrum of voiced excitation contains discrete components at harmonics of the pitch frequency. For unvoiced sounds, the excitation is due to the turbulence generated by air flow through partially opened glottis. The excitation tends to be random in nature, with a flat, continuous spectrum. Mixed excitation is also possible for the class of sounds known as voiced fricatives, in which turbulent excitation is amplitude modulated periodically by the vibration of the vocal cords. Anatomical and physiological properties of vocal cords or glottis which is considered as source of the speech production mechanism varies for different speakers. When different speakers try to produce the same sound, their vocal tracts

TH-1853_02610201

are positioned in a similar manner. The actual shapes will be different due to the differences in the anatomical structure of the vocal tract. Variability due to differences in the anatomical structure of the vocal tract and the source corresponds to the speaker part of the speech signal. Researchers have tried to extract and use these information from the speech signal for solving the problem of speaker recognition.

Speaker recognition can be performed as speaker identification, speaker verification, speaker spotting and speaker detection. Out of the above four categories of speaker recognition, speaker identification and speaker verification are most widely used [3]. Speaker identification system identifies a speaker from a set of reference speakers. In automatic speaker identification, there is no prior identity claim and the system decides who the person is, what group the person is a member of, or whether the person is unknown. Speaker identification can be performed as open-set identification or closed-set identification. In open-set identification, the possibility exists that the unknown voice token does not belong to any of the reference speakers. In closed-set identification, the voice token belongs to one of the speakers in the reference set. Validity of a speaker can be verified by Speaker verification system [4–8]. The literature abounds with different terms of speaker verification, including voice verification, speaker authentication, voice authentication, talker authentication and talker verification.

The speaker recognition system consists of two major problems. First problem is the extraction of a set of speaker dependent features. Second problem is the choice of a suitable model or classification system. Figure 1.1 shows a typical speaker recognition (SR) system. Mainly, speaker recognition system consists two phases. They are training phase and testing phase. Training phase creates the reference speaker model using training utterances for each speaker. In testing phase, the testing utterances are used for recognition of speaker. Using a similarity measure between the test utterances and the reference speaker's model, the decision for speaker recognition is taken. Figure 1.2 shows speaker identification system and speaker verification system. Speakers' reference models are same for both the cases, but the testing method is different. In speaker identification, testing utterance is used to find a close speaker from reference speaker set. In this case, maximum matched speaker's identity is selected as the recognized speaker. In speaker verification, identity of the speaker is mentioned with the testing voice token. So the testing voice token is verified only with that particular speaker's voice. The verification is done with a threshold. This threshold can be fixed individually or globally.

When the speaker is cooperative, the system may know what the speaker is supposed to say and better performance may be achieved. When the speaker is uncooperative or unaware, then the system will be handicapped by the lack of this knowledge. There are two different modes used in speaker

TH-1853_02610201

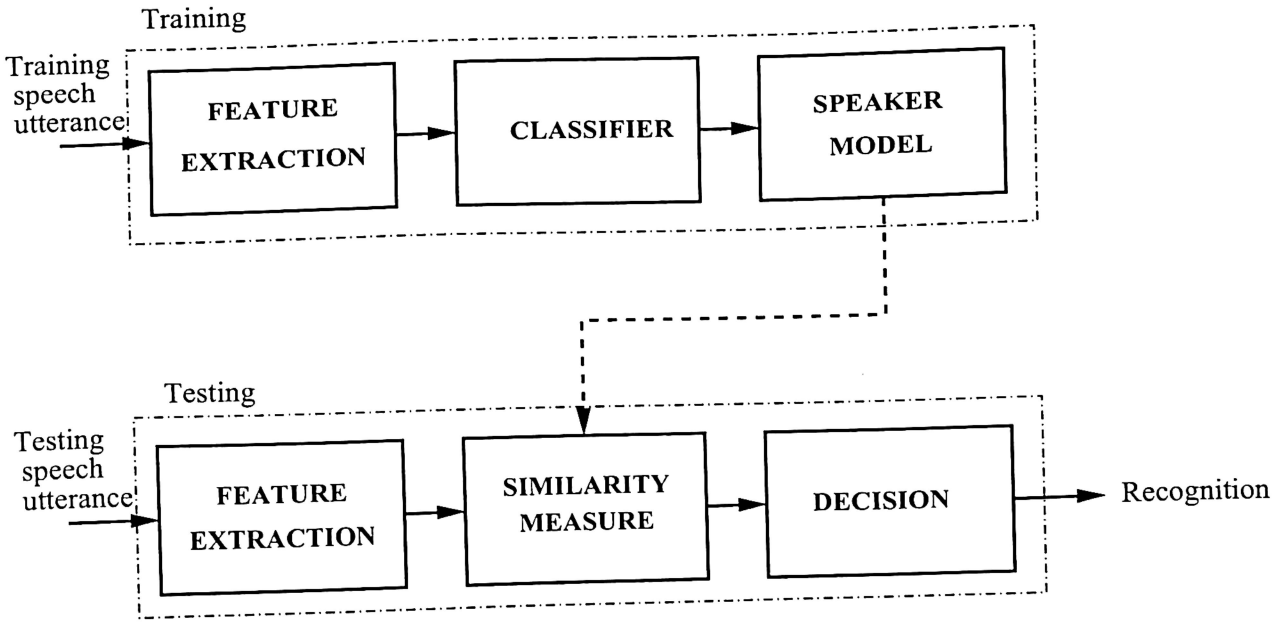


Figure 1.1: Block diagram of speaker recognition system

recognition, which correspond to two different kinds of applications. One is Text-dependent speaker recognition and the other is Text-independent speaker recognition.

1.1.1 Text-dependent speaker recognition

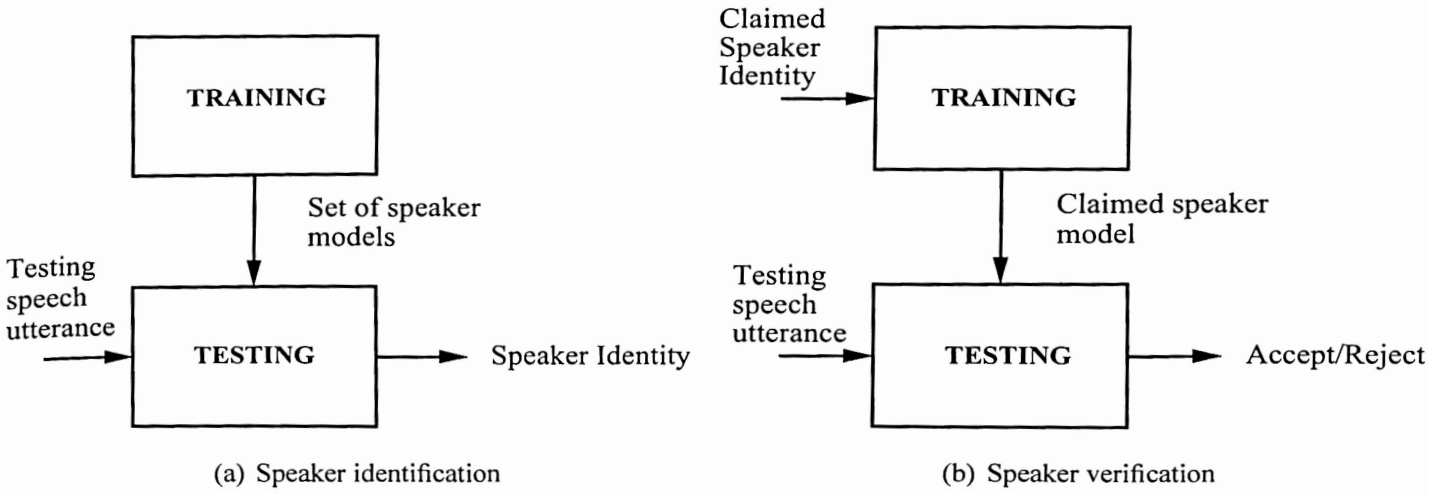
A speaker recognition system is called "Text-dependent" if it knows what the speaker is supposed to say. In the typical text-dependent recognition scenario, the speaker will either say a predefined utterance or will be prompted to say an utterance by the system. Text-dependent speaker recognition methods can be classified into DTW (dynamic time warping) or HMM (Hidden Markov model) based methods [9]. In either case, the target speaker will have explicitly spoken these words during an enrollment session.

This explicit knowledge can be used to good effect in building detailed dynamic models of the speaker. These models can be word-specific and phone-specific and thus can calibrate the target speaker with good accuracy. With nominal speaker co-operation, text-dependent recognition improves recognition performance while at the same time minimizing the amount of speech data required. Sivakumaran et. al have analyzed the sub-band based text-dependent speaker verification [10].

1.1.2 Text-independent speaker recognition

A speaker recognition system is called "Text-independent" if it does not have the knowledge of what the speaker is saying. This forces the technology to deal with whatever speech data happens to be

TH-1853_02610201



(a) Speaker identification (b) Speaker verification

Figure 1.2: Speaker recognition systems, (a) speaker identification system and (b) speaker verification system.

available, both for training and for recognition decision. Because of this, text-independent systems typically exhibit lower performance than text-dependent systems. In this case, to learn the voice characteristics of speaker, phonetically well balanced text is needed for training. Long-term-statistics-based methods [11], [12] , VQ-based methods [13] and Ergodic-HMM based methods [14] are used for text-independent speaker recognition.

There is significant research advantage to work on text-independent recognition, beyond just the use of large amounts of speech data. This is, namely, the general absence of idiosyncrasies in the definition of the task and the specification of speech data. This allows simple and general task, which in turn supports much broader collaboration and sharing of results than in text-dependent recognition. As a result of this advantage, interest in text-independent recognition has risen significantly. Reynolds et. al proposed text-independent speaker verification using adapted Gaussian Mixture model [15].

1.2 Major Tasks of Speaker Recognition Systems

Feature extraction, classifiers, similarity matching and decision logic are the major tasks in speaker recognition systems. Speech features reflect the characteristics of the speakers. Classifiers are used for making reference model of each speaker using training feature vectors. These classifiers help accumulate the speaker information into a particular area in feature space. In testing phase, these speaker models are used for similarity matching in terms of score. Decision of speaker recognition is taken based on similarity matching score.

TH-1853_02610201

1.2.1 Feature extraction

Features reflect the characteristics of the speech production system. Speaker identity is correlated with the physiological and behavioral characteristics of the speech production system for each speaker. These characteristics exist both in the spectral envelope (vocal tract characteristics) and in the supra-segmental features (voice source characteristics) of speech signal [9]. Although it is impossible to separate these characteristics, and it is difficult to measure them explicitly, many characteristics are captured implicitly by various signal measurements. Signal measurements such as short-term and long-term spectra and overall energy are easy to estimate. These measurements provide the means for effectively discriminating the speakers. The speech features are extracted for training and testing as voice print representation of a speaker.

Fundamental frequency is used as an important feature for recognition of speakers [4]. Reynolds [16] has experimented with telephone-speech database for evaluation of performance of different features such as Mel-frequency, linear-frequency filter bank cepstral coefficients, LP cepstral coefficients and perceptual linear prediction (PLP) cepstral coefficients. Best results are reported for higher order linear prediction (LP) features and band-limited filter bank features. Different types of features are investigated and the LP-derived cepstral coefficients are usually found to yield the best performance [17]. Campbell [6] has provided a description of different features such as reflection coefficients, log area ratios, arc-sin reflection coefficients, line spectrum pair (LSP) frequencies and mel-frequency cepstral coefficients (MFCC). Evaluation of relative performances of different features such as LP cepstrum, autoregressive moving average (ARMA) cepstrum, log area ratios and reflection coefficients is carried out under different noise conditions [8]. Guruprasad et al have compared the performance of speaker recognition systems based on cepstral coefficients with the systems based on difference cepstral coefficients [18]. LP derived cepstral coefficients have been used for a speaker-specific mapping for text-independent speaker recognition [19]. A set of new dynamic features like delta cepstral energy (DCE) and delta-delta cepstral energy compactly represent the information [20]. The pitch adjustment operation attempted to better align the frame boundaries to the raw signal, thus reducing the spectral estimation artifacts [21]. All these results show the wide acceptance of cepstral coefficients for speaker recognition applications.

The LP-cepstral coefficients distinguish one speaker from another by detecting the differences in their glottis and the vocal tract which are reflected in the LP-spectrum. The speaker specific sound information of phonemes, long vowels and nasals are captured by cepstral features. However, the

TH-1853_02610201

भारतीय प्रौद्योगिकी संस्थान गुवाहाटी • Indian Institute of Technology Guwahati

differences between the speakers can be also observed in the frequency domain as shown by Liu at el. [22]. The comparative tests have shown the advantages of the line spectrum pairs (LSP) over the LP-cepstral features. Imperl et al [23] analyzed the spectral (Hildebrand-Prony) information captured by harmonic features for speaker identification. Prony harmonic features are found to be more effective than LPCC features especially they are superior in case of speaker identification. This shows that the harmonic features are more influenced than the LP-based features.

Umesh et al warped the mel scale according to the experimental results which show that the scale-factor relating the formant frequencies of different speakers increases with decreasing values of formant frequencies [24]. Different models, each with few parameters, are used for the experimentally obtained mel scale and they are analyzed for speaker recognition [24], [25], [26]. To generate features for speaker recognition, autoregressive model is used for the residual signals which are produced from the vocal tract inverse filtering. These features are compared with the conventional MFCC features [27]. Improved performance of MFCC has been shown with residual phase for speaker verification [28]. Chen et al have combined the residual and the prosodic features with the acoustic features for speaker recognition [29]. Kim et al have used pitch synchronous MFCC features for speaker recognition [30].

Kaja et al [31] have proposed combination of prosodic and lexical features for speaker recognition. Dynamic features (velocity and acceleration) are used to improve the speaker recognition [32]. The system uses an auditory-based front-end processing and incorporates new algorithms for the extraction and manipulation of the acoustic phonetic features that proved to be rich in their information content [33]. Yegnanarayana et al have investigated the source (LP residual) and system features (weighted linear prediction cepstral coefficients) for speaker recognition using AANN [34]. Sivakumaran has focused on the spectral representation of the sub-band cepstrum in relation to that of the full-band cepstrum [35]. Wavelet based decomposition of cepstral features are investigated for inter-speaker separation [36], [37]. Articulatory features (AF) are derived from spectral features for telephone-based speaker verification [38]. A supplementary feature is investigated for speaker verification in constrained room acoustics [39]. In long-term analysis, trends in pitch variation are believed to appear at different time-scales, such as microprosody, accent, phrase and discourse levels, making wavelet analysis of the pitch contour a suitable choice for investigating the corresponding pitch patterns for speaker identification [40]. Researchers at Speech Technology and Research Laboratory (Menlo Park, CA, USA.) have used the pitch frequency (Pitch prosodics) as an additional feature to model the static and the dynamic pitch contour information [3]. The speaker independent information is suppressed by mapping technique [41], [19]. Suprasegmental features are also well known features for speaker recognition [42].

TH-1853_02610201

Integration of different feature sets for speaker recognition has been proposed by Chagnolleau et al [43]. This approach, named as time-frequency principal components (TFPC), seeks to find out the time dependent spectral vectors from the speech signal and these are used for speaker recognition.

1.2.2 Classifiers

In speaker recognition systems, classifiers are used for building speaker reference models using training feature vectors. These classifiers help concise the data which represent speaker information. The speaker models can be categorized into models with supervised training or models with unsupervised training [44]. In supervised training, the speaker models are trained together with separate indices. In unsupervised training, the speaker models are trained individually. Advantage of supervised method over unsupervised method is that the speaker model captures better information between a particular speaker and other speakers in the reference set. Disadvantages of supervised method over unsupervised method are more computations and large data. Examples of unsupervised methods are Nearest Neighbor(NN), Vector Quantization(VQ), Gaussian Mixture Model(GMM), Hidden Markov models(HMM). Examples of supervised methods are Multilayer Perceptrons (MLP), Radial basis functions (RBF) and decision trees and Neutral tree networks(NTN) [45]. Hidden Markov Model [46], Neural Networks [47] and Gaussian Mixture Model [48] are widely used classifiers for speaker recognition systems.

Nearest neighbor (NN) and Vector quantization (VQ)

The NN, or more general kNN, method can be used to estimate the probability density function (PDF) of the data within a class. The simplest KNN classifier is the NN classifier (kNN with k = l), which operates as follows. All training data, i.e., each feature vector along with its corresponding label, are stored. The distances between a given test vector and all training data are computed. The test vector is then assigned the label of the training vector with the smallest distance, i.e., the nearest neighbor [49]. More computation and data storage space is required for this NN method. To decrease the number of reference vectors, Vector Quantization (VQ) method has been preferred. A popular iterative approach for computing the clusters is called the K-means algorithm. Here, K refers to the number of classes or modes. Mostly, K-means algorithm referred as VQ method. Linde-Buzo- Gray (LBG) method [50], the Learning Vector Quantization (LVQ) method [51] and Group Vector Quantization (GVQ) method [52] are common VQ methods. VQ method is successfully used for speaker recognition [53].

Dynamic time warping (DTW)

TH-1853_02610201

Dynamic time warping (DTW) is a template matching technique which has been successfully used for text-dependent speaker recognition [17, 53–57]. The time function of the vector X_t in the voiced portion is used directly for speaker recognition. An input utterance is brought into time registration with a reference template to calculate the distance between them. When the time duration varies between utterances, nonlinear time warping method using a dynamic programming technique [58] is used.

Gaussian mixture model (GMM)

The GMM method is a statistical technique based on log likelihood function [59], [60]. The density of log likelihood function is a linear combination of mixture of multivariate Gaussian densities. Multivariate Gaussian depends upon the mean and covariance. If we consider acoustic class like unimodal Gaussian representing phonetic events such as vowels, nasals or fricatives, speaker dependent vocal tract configuration is reflected by these events [61]. In GMM, these event mixtures are represented by Gaussian mixtures. In recent years, it is common to represent speakers with the Gaussian mixture model (GMM) [62], [63], which can be considered as one single-state HMM. The linear combination of Gaussian functions is capable of representing a large class of sample distributions of speakers utterances [61]. Well-known technique of constrained maximum likelihood linear regression (MLLR) is extended to compute a projection (instead of a full rank transformation) on the feature vectors of the adapted data [64].

Hidden markov model (HMM)

The HMM is a time dependent classifier. Here, time-dependent parameters are observation symbols and continuous probability measures for discrete HMM and continuous HMM respectively. Observation symbols are created by VQ codebook labels. Continuous probability measures are created using GMM. The states are not directly observable and the observations are probabilistic functions of the state [65, 66]. Main assumption of HMM is the current state depends on the previous state. In training phase, state transition probability distribution, observation symbol probability distribution and initial state probabilities are estimated for each speaker as a speaker model. The probability of observations for given speaker model is calculated for speaker recognition [67–69].

Continuous ergodic HMMs identified speakers much more accurately than discrete ergodic HMMs [66]. Continuous HMMs are as resistant to session-to-session variations in speech and to those due to different utterance rates as the VQ distortion based method. However, when less data is available, the VQ distortion based method is more robust than a continuous HMM based method. With continuous

TH-1853_02610201

ergodic HMMs, the speaker identification rates are strongly correlated with the total number of mixtures, irrespective of the number of states. This means that the information about transitions between different states do not contribute to text independent speaker recognition.

Neural Networks

The neural approach out-performs the codebook system for small model sizes. In both training and recognition modes, the effect of the textual content must be considered. The DTW and HMM approaches are highly text-dependent and are unlikely to recognize a speaker unless the appropriate text is spoken. A high performance speaker recognition system should highlight speech characteristics which are ideally unique to a given person, and of course both the training data and the testing sequences must contain such sounds. It follows that, the training data should contain as diverse a range of sounds as is practicable, personal models that highlight speaker-dependent sounds should be used, and a positive recognition decision should be made only if sufficient discriminating sounds are spoken in the test utterance [70].

One limitation of HMMs is that they generally require more data than other classifiers at the training stage. The DTW method uses a distortion measurement between the extracted features of a test utterance and those stored for that speaker during training. Hence, it provides an intraspeaker measure. Additionally, it uses temporal information, which is not used in the NTN. Since the information provided by the NTN and DTW classifiers are somewhat independent, they can be combined to yield a powerful classification system for text-dependent speaker verification [71]. The neural networks are trained to model differences among known speakers, which allows a small number of parameters (more efficient to train and test) and can model the decision function for discriminating speakers within a known set [72]. Mak et al incorporates full covariance matrices into the radial basis function (RBF) networks (Elliptical basis function) and uses the expectation-maximization (EM) algorithm to estimate the basis function parameters for speaker verification [73].

1.2.3 Similarity matching

Matching techniques generally utilize the uniqueness in the speakers' feature vector distributions for discriminating the speakers. Generally, there will be confusion between speakers due to the reasons such as, intra-speaker variability, inter-speaker confusability and environmental degradation. The matching technique should take into account these factors for discriminating the speakers. Classifiers

TH-1853_02610201

are used for identifying a speaker based on template matching or maximum log-likelihood value. In template matching, the feature vectors collected from the training set are stored as reference and the feature vectors collected from test data are compared with the reference using a suitable distance measure. Probabilistic modeling involves modeling of speakers by probability distribution, and deriving the classification decision based on the probability or likelihood. Probabilistic modeling is used mainly in the Text-independent systems.

Template matching

Template matching is generally used in text-dependent speaker recognition systems. The template model and its corresponding distance measure is perhaps the most intuitive method. The template method can be dependent or independent of time. An example of a time-independent template model is VQ modeling [6]. All temporal variations are ignored in this model and global averages (e.g., centroids) are used. A time dependent template model is more complicated because it must accommodate variability in the human speaking rate. Initially, the Euclidean distance is calculated for distance measurement. Then the Mahalanobis and weighted distances are selected for better recognition [74]. The template representation of a small portion or frame of speech using N features or parameters can be viewed as an N-dimensional vector (or point in N-dimensional feature-space) [75]. The similarity between two templates can be viewed as inversely proportional to their separation distance in N-dimensional space. One standard measure is a quadratic distance. It is calculated between two N-dimensional templates x and y.

$$d(x, y) = (x - y)^T W^{-1} (x - y) \tag{1.2.1}$$

where W is a positive-definite matrix which allows different weighting for individual features of the template, depending on their utility in identifying the speakers in the feature space. The common Euclidean distance sets W to be the identity matrix, whereas the general Mahalanobis distance sets W to be the autocovariance matrix corresponding to the reference vector. Individual matrices could be used for each reference template, but an average W is more commonly used without significant accuracy loss, because it is difficult to obtain accurate estimates for speaker-individual W from limited training data and it is efficient to use one W matrix.

The most popular method to compensate for speaking-rate variability in template-based systems is known as DTW [2]. At the end of the time warping, the accumulated distance is the basis of the match score. This method accounts for the variation over time (trajectories) of parameters corresponding to the dynamic configuration of the articulators and vocal tract [6].

TH-1853_02610201

Probabilistic Modeling of Speakers

Probabilistic modeling of speakers refers to modeling speakers by probability distributions rather than by average features and to basing classification decisions on probabilities or likelihoods rather than distances to average features [75]. Assuming that the distributions for speakers are known and have continuous densities p_i then the likelihood that a feature x is generated by the i^{th} speaker is $p_i(x)$. Using Bayes' rule, the probability that the speaker is the i^{th} speaker is

$$P(speaker = i/x) = \frac{p_i(x)P_i}{p(x)} \tag{1.2.2}$$

where P_i is the prior probability that the utterance came from the i^{th} speaker, and $p(x)$ is the probability of the feature x occurring from any speaker. Typically the prior probabilities for the speakers are assumed equal. The term $p(x)$ is the average of the speaker densities,

$$p(x) = \sum_{i=1}^{I} p_i(x)P_i \tag{1.2.3}$$

where I is the number of speakers. Note that $p(x)$ is the same for all speakers and if the prior probabilities are equal, then the speaker to choose will simply depend on which speaker has the highest likelihood. This will be the most probable speaker given the observed feature, and is known to result in the minimum error strategy.

Background speakers

Likelihood ratio or cohort normalized scoring has been shown to be effective for improving the performance of speaker verification systems [76]. An important problem in this connection is the establishment of principles for constructing speaker background or cohort models which provide the most effective normalized scores. These include individual speaker models, models constructed from the pooled utterances of different numbers of speakers, models selected on the basis of similarity with customer models, models constructed from random selections of speakers, and models constructed from databases recorded under different conditions than the customer models. The results of experiments show that pooled models based on similarity to the reference speaker perform better than individual cohort models from the same set of speakers.

The inter speaker confusability can be reduced by background speakers [16]. Two issues that arise with the use of background speakers are the selection of the speakers and the number of speakers to use. Intuitively, the background speakers should be selected to represent the population of expected

TH-1853_02610201

imposters, which is in general application specific. In some scenarios, it may be assumed that imposters will attempt to gain access only from similar sounding or at least same-sex speakers (dedicated imposters).

In a telephone based application accessible by a larger cross-section of potential imposters, on the other hand, the imposters may sound very dissimilar to the users they attack (casual imposters); for example a male imposter claiming to be a female user. Previous systems have relied on selecting background speakers whose models are closest or most competitive for each enrolled speaker (termed as the ratio-sets or cohorts). This occurs because the dissimilar voice is not modeled well by either the numerator or denominator of the likelihood ratio. The test is based on tails of distributions of the speaker models giving rise to unreliable values. The experiments can be examined both for the same-sex and mixed-sex imposter situations. Ideally, the number of background speakers should be as large as possible to better model the imposter population, but practical considerations of computation and storage dictate a small set of background speakers.

In the verification experiments, limited number of background speakers are selected. The limited size was motivated by real-time computation considerations and the desire to set a constant experimental test. For a verification experiment on a given database, each speaker is used as a claimant with the remaining speakers (excluding the claimants background speakers) acting as imposters and rotating through all speakers. Large background speaker sets decrease the number of imposter tests. This paradigm allows the maximum use of the speakers for a large number of tests but makes it difficult to compare systems using different size background speaker sets. An alternative approach is to partition the database speakers into separate claimant and imposter groups with background speakers selected only from the claimant group.

1.2.4 Decision logic

Decision logic is mainly used to arrive at a recognition decision based on the result of matching. Matching generally gives a score representing how well the test feature vectors match with the reference. If the score is high then, the match is good. A simple decision logic is to fix a threshold for the scores. A speaker will be accepted if the score is above the preset threshold. The decision logic can also be complex involving combining the evidences from several matching scores. Combining evidences can be as simple as addition of the scores or a sophisticated method of using constraint satisfaction model [28].

The verification problem is traditionally formulated as "statistical hypothesis testing", where two

TH-1853_02610201

complementary hypothesis are used, each corresponding to one of the yes and no answers accordingly [77]. The Bayesian approach to verification involves the evaluation of a quantity called the Bayes factor which involves evaluation of several Bayesian predictive densities against a critical threshold. Standard overall performance evaluation of speaker recognition system is shown in terms of detection error trade-off (DET) of missed detections and false alarms for each system and the effects on performance of training condition, test segment duration, the speakers' sex and the match or mismatch of training and test handsets [3]. Frame likelihoods are transformed into new scores according to some Non-linear frame likelihood transformation [78]. To reduce the equal error rate, score normalization techniques are introduced.

Score normalization

Score normalization has been introduced explicitly to cope with score variability and to make speaker-independent decision threshold tuning easier [7]. T-normalization and Z-normalization techniques reduce the equal error rate in speaker verification. Test utterance normalization (T-normalization) is performed on the raw confidence score in order to transform the scores into a similar range. In Z-normalization, a speaker model is tested against a set of speech signals produced by some imposter, resulting in an imposter similarity score distribution. Speaker-dependent mean and variance-normalization parameters are estimated from this distribution and are applied on similarity scores yielded by the speaker verification system [79]. One of the advantages of Z-normalization is that the estimation of the normalization parameters can be performed offline during speaker model training. The speaker recognition system scores will have confusability in noisy stressful environment. These confusability factors are more challengeable in speaker recognition system.

1.3 Speaker Recognition and Stressed Speech

A person's speech exhibits distinctive characteristics which indicate the identity of the speaker. Speaker dependent speech features capture these characteristics. Several speech features, discussed in section 1.2.1, are used for this purpose. In literature, features based on spectral information are widely used for speaker recognition. Features which are commonly used for speaker identification (SI) include cepstral features [80], Mel-frequency cepstral coefficients [6], Harmonic features [23], sinusoidal model features [81], [82] and suprasegmental and source features [42].

TH-1853_02610201

Due to various reasons, the above features from the same speaker show wide range of variability. While it is the speaker's voice that provides the recognition capability, it's also the speaker's variability that makes the speaker recognition problem so difficult. An explanation for why the speaker's variability is such a vexing problem is that the use of speech, unlike finger prints or hand prints or retinal patterns, is to a very large degree a result of what the person does, rather than who is the person. Speech is performing art and each performance is unique. Some of the sources of variability are not within the speaker's control. There are other factors, beyond speaker's variability, that present a challenge to speaker recognition technology [3]. These deal with problems like, acoustical signal intact from the speaker to the recognition system, mimic, closely related members like identical twins , microphone and electrical transmission [83].

One of the major causes of variability in a speaker's voice can be attributed to stressed condition. Stressed speech [84], [85], [86] is defined as the speech produced under any condition that causes the speaker to vary speech production from neutral condition. The causes of perceptually induced stress include emotion and environmental factors such as noise. In stressed condition, the characteristics of speech signal are different from that of normal or Neutral condition. Due to changes in speech signal characteristics, the performance of speaker recognition system may degrade under stressed speech conditions. There will be reduction in speaker recognition performance if speech signal recorded in one stressed condition is used for training and speech recorded in another stressed condition is used for testing and vice-versa. Normally, during the training sessions, the signals are expected to be recorded in Neutral emotion condition. As the speech signal used for recognition is recorded in a different condition, this speech signal may be different from that of the Neutral emotion condition. At the time of recognition, the stress condition will depend not only on the environment but also on the person's state of mind. As result, the testing speech signal may be considered as that from a different stress condition compared to the speech signals during the training session. During the last two decades, though some research publications in the area of recognition of stressed speech are published, there is little effort for analysis and recognition of speakers under stressed condition. Following subsection presents a review of works carried out in the area of stressed speech.

1.3.1 Analysis and recognition of stressed speech

Most of the research efforts in stressed speech are aimed at the analysis of stressed speech signal for possible improvement in the performance of speech recognition systems [87]- [88]. Primary objective

TH-1853_02610201

of various works presented in literature is the recognition of emotions from stressed speech signal. Conventional speech features, which are widely used in various speech processing applications, are investigated to model stressed speech signals under different emotions. Cowie *et al.* [89] have presented a review of speech features such as pitch frequency, intensity, duration and spectral features which are used by various authors for recognition of emotions. Pitch contours, syllabic durations and intensities are used for recognition [90] and synthesis [91] of emotional speech. High frequency energy and articulation rate with pitch are investigated for possible correlation with the emotions [92]. Dimitrios *et al.* [93] have discussed the importance of pitch, formants, vocal tract cross-section areas, mel-frequency cepstral coefficients, Teager energy operator based features, intensity and speech rate for recognition of emotions. Linear Prediction (LP) and cepstral features are also tested for analysis of stressed speech signals [94]. Fast Fourier Transform (FFT) based linear short time Log Frequency Power Coefficients (LFPC) and Teager Energy Operator (TEO) based nonlinear LFPC features are used for recognition of stressed speech [95], [96]. Though Intensity and duration features are shown to be effective in recognition of stressed speech, they are seldom used in speech and speaker recognition algorithms. Frequency, amplitude and phase features of the sinusoidal model are analyzed and used as input features for recognition of different stress classes [85]- [97]. Wavelet packet parameters and sub-band based cepstral parameters are shown to outperform traditional mel-frequency cepstral coefficients for monophone stressed speech recognition [98].

Studies on speech signals produced under different stressed and noisy conditions have helped in improving speech recognition algorithms. Different features are tested for speech recognition with the aim of finding robust features for this purpose [99]. Hansen et al [100] have proposed a feature enhancement and source equalization technique for improving the performance of automatic speech recognition for different noisy and stressed conditions. A real-time speech recognition system [84] has been designed for recognition of speech in noisy stressful and Lombard effect environments. In speech recognition applications, the FFT based features are observed to be more robust to noise compared to the LP based features [101], [99] and [102], whereas it is less robust to different stressed conditions in a speech signal [101]. This suggests that the characterization of different stressed conditions present in a speech signal will be possible through FFT based features. Hence, further investigation into FFT based features is required for their use as relayers of different stressed conditions in a speech signal. Furthermore, Vector-Quantization (VQ) method, which has been widely used as a classifier [2] [103] for speech recognition, requires more attention for classification of different stress classes in a speech signal.

TH-1853_02610201

1.4 Scope of the Present Work

Though several works are reported in literature on analysis and recognition of stressed speech, there is little effort in the direction of speaker recognition under stressed condition. All most all the works in the area of speaker recognition focus on speech under Neutral condition. Recently Shahin [104] has shown that the deterioration in performance of a speaker recognition system for shouted talking condition can be controlled by using circular Hidden Markov Model. Above facts suggests that a detailed investigation is required for the performance of speaker recognition algorithms under stressed condition. It is necessary to quantify the degradation of the performance of the speaker recognition algorithms under different stressed conditions.

It is expected that the features which carry speaker dependent information are effected due to stress condition. So, it will be useful to analyze the speaker dependent features under stressed conditions. Most of the speech features are used for recognition of stressed classes or emotions in a speech signal. These features are common in speech and speaker recognition applications. It would be interesting to find out the relative performance of these features for speaker recognition under stressed conditions. Different speech features are tested for characterization of stressed conditions with various stressed speech databases. Researchers have also used different classifiers, such as Hidden Markov Model (HMM), Neural Networks [105], etc., for evaluation of performances of stressed speech features for recognition of different stress classes. Though these classifiers show good recognition results for classification of different stressed conditions, they do not provide meaningful insight into the capabilities of various speech features in the characterization of different stressed conditions in speech signals. In cepstral analysis, the excitation source information which appear in terms of high-time component of the cepstrum is mostly ignored in speaker recognition applications [28]. Both source and vocal tract are effected during stressed conditions. Analysis of features from the point of view of source and vocal tract separately may help establishing the fact whether the source or vocal tract is more effected by the stressed condition.

Cepstral features such as mel-frequency cepstral coefficients (MFCC) [28], cepstral coefficients [106] and linear prediction (LP) cestral coefficients are used by various authors for speaker recognition. MFCC features, which are the most effective among the three, are estimated from the spectral energies from different mel-frequency bands. In this way, the information associated with the spectral peaks (amplitudes and bandwidths) are not explicitly used. LP coefficients and LP derived features [2], such as reflection coefficients [107], arcsin reflection coefficients [6], log-area ratios [58], are also used

TH-1853_02610201

for speaker recognition. All the above features are spectral domain features extracted from a preemphasized speech signal. Preemphasis enhances the vocal tract information. The formants appear as spectral peaks in a preemphasized speech signal. Amplitudes, frequencies and bandwidths of the formants have significant speaker information. The spectral domain features used for speaker recognition do not use these information explicitly. Authors have shown that performance of speaker recognition can be improved if different features can be used by combining them in a suitable manner [43]. A new set of features which can eliminate the above limitations will be useful towards improvement of performance of speaker recognition systems.

Various compensation techniques are used for improvement in speech recognition under stressed condition. Hansen has proposed parameter based compensation techniques for recognition improvements [48, 84, 94, 105, 108–110]. The features are enhanced to address noise and an adaptive mel-cepstral compensation algorithm is used to equalize the impact of stress. Roe [111] has proposed a traditional template based system that incorporated both vector quantized and replaced by the closest spectral codeword in the VQ codebook for subsequent distortion calculation. The spectral distortion induced by unusual speaking efforts could be compensated by simple linear transformation of the cepstrum [112]. The basic assumption of this technique was that spectral distortion induced by unusual speaking efforts could be compensated by simple linear transformation of the cepstrum. Although this assumption may be unduly optimistic, it was observed that the statistics of cepstral vectors did display some systematic modification in various speaking styles. The probability of spectral compensation in a stochastic model based recognizer was found to be effective in dealing with various articulation effects. There is almost negligible effort for analysis and improvement of performance of speaker recognition systems under stressed conditions. Feature compensation techniques may be experimented for enhancement of performance of speaker recognition systems.

The work presented in this thesis is broadly divided into three categories. First, the performance of the speaker recognition systems under stressed conditions is evaluated. The effects of different stressed conditions on speech features are analyzed and their relative performance for speaker recognition is evaluated. Second, sinusoidal model is tested for speaker recognition. Sinusoidal model has three features. They are amplitude, frequency and phase. Assuming that the information in these three features may be different, these features may be combined for improvement in speaker recognition under stressed condition. Finally, compensation techniques are proposed and evaluated for improvement of performance of speaker recognition systems. Compensators are aimed at reducing the stress effects.

TH-1853_02610201

features. They are amplitude, frequency and phase. Assuming that the information in these three features may be different, these features may be combined for improvement in speaker recognition under stressed condition. Finally, compensation techniques are proposed and evaluated for improvement of performance of speaker recognition systems. Compensators are aimed at reducing the stress effects.

1.5 Organization of the thesis

The organization of the thesis is as follows.

Chapter 2 presents a review of speech features and classifiers used for speaker recognition. Mel-frequency cepstral coefficients (MFCC), Linear prediction (LP) coefficients and LP derived features such as reflection coefficients, log-area-ratios, arcsin reflection coefficients and cepstral coefficients are investigated for their relative performance in speaker recognition. Section 2.1 presents a brief review of the features. Various statistical feature evaluation techniques used in this work are discussed in Section 2.2. Two classifiers, Vector Quantization (VQ) and Gaussian Mixure Model (GMM), are presented in Section 2.3. This section also discusses speaker recognition system and its performance with different features. Section 2.4 presents speaker recognition performance of different features under different stressed conditions and discusses the motivation for this work.

In Chapter 3, we present the analysis and evaluation of six speech features, MFCC, LP coefficients, LPCC, RC, ARC and LAR, under stressed condition. Two classifiers, vector quantization (VQ) classifier and Gaussian Mixture Model (GMM) are used to evaluate the recognition results under Neutral and stressed conditions. In Section 3.1, we present the statistical evaluation of the features. Probability density function (pdf) characteristics, F-ratio test and Kolmogorov-Smirnov (K-S) test are used for evaluation of speaker discrimination capability of the features under stressed conditions. Section 3.2 presents the analysis of the features based on vector quantization (VQ) and Gaussian mixture model (GMM) based classifiers. The relative ranking of the features based on their performance in speaker recognition under stressed condition is evaluated in Section 3.3.

Chapter 4 presents speaker recognition using sinusoidal model features. Frequency, Amplitude and Phase features of the sinusoidal model are analyzed and used as input features to a speaker recognition system. Speaker recognition performance of the three sinusoidal model features is evaluated under stressed speech conditions. In Section 4.1, sinusoidal modeling of speech is presented. Statistical analysis of sinusoidal model features is discussed in Section 4.2. Speaker recognition performance of the sinusoidal model features under stressed speech condition is discussed in Section 4.3. A novel approach

TH-1853_02610201

speech signal. Four compensation techniques are proposed and tested for improvement of speaker recognition under stressed condition. In Section 5.1, we propose speaker and stress information based compensation (SSIC) technique. Section 5.2 presents compensation by removal of stressed vector (CRSV). Section 5.3 presents combination of MFCC and SAF (CMSA) as a compensator. Cepstral mean normalization (CMN) is used for compensation in Section 5.4. Finally, we present selection of compensator by stress recognition in Section 5.5.

Finally, we conclude in chapter 6 by summarizing the contributions of the thesis and noting some directions for further investigations in these areas.

TH-1853_02610201

Chapter 2

Review of Speech Features and Speaker Recognition

Features based on spectral information are widely used for speaker recognition. Features which are commonly used for speaker recognition include cepstral features [80], mel-frequency cepstral coefficients [6] and linear prediction (LP) cepstral coefficients. Linear prediction (LP) coefficients and LP derived features [2], such as reflection coefficients [107], arcsin reflection coefficients [6], log-area ratios [58], are also used for speaker recognition. Researchers have also used various classifiers for speaker recognition. Vector Quantization (VQ) [2], Hidden Markov Model (HMM) [46], Neural Networks [47] and Gaussian Mixture Model (GMM) [48] are widely used classifiers for speaker recognition. Performance of a speaker recognition system not only depends upon the speech features and the classifiers but also on the environment. For the same speaker, in a noisy environment the speech features may vary from the speech features extracted from speech signal recorded in a silent environment. In a noisy environment, the speaker is stressed. This causes the speaker to vary his speech production from normal or neutral condition. Performance of speaker recognition system will degrade if the training speech data and the testing speech data are collected in different stress conditions.

In this chapter, we present a review of speech features and classifiers used for speaker recognition. Mel-frequency cepstral coefficients (MFCC), Linear prediction (LP) coefficients and LP derived features such as reflection coefficients, log-area-ratios, arcsin reflection coefficients and cepstral coefficients are investigated for their relative performance in speaker recognition. In addition to Vector Quantization (VQ), Gaussian Mixure Model (GMM) classifiers, statistical feature evaluation techniques, probability density characteristic, F-ratio test, Multi-dimensional F-ratio test or divergence, KS-test, are used for evaluation of features. Section 2.1 presents a brief review of the features. Various

TH-1853_02610201

statistical feature evaluation techniques used in this work are discussed in Section 2.2. Two classifiers, Vector Quantization (VQ) and Gaussian Mixure Model (GMM), are presented in Section 2.3. This section also discusses speaker recognition system and its performance with different features. Section 2.4 presents speaker recognition performance of different features under different stressed conditions and discusses the motivation for this work.

Speaker Databases

In this work, two speaker databases, YOHO database and simulated and actual stress (SUSAS) database, are used. The YOHO database was designed to support text-dependent speaker verification research such as is used in secure access technology [16]. It has a well defined train/test scenario in which each speaker has four enrollment sessions where he/she is prompted to read a series of 24 combination- lock phrases. Each phrase is a sequence of three two-digit numbers (e.g., 35-72-41, pronounced thirty-five seventy-two forty-one). There are 10 verification trials per speaker consisting of four phrases per trial. The vocabulary consists of 56 two-digit numbers ranging from 21 to 97. The speech was collected in an office environment using a telephone handset connected to a workstation. Thus, the data has a telephone bandwidth of 3.1 kHz (300-3400 Hz), but no telephone transmission degradations.

The second speech database used in this work is a stressed speech database which is popularly known as speech under simulated and actual stress (SUSAS) database [98]. From this database, four stress classes or emotions, Anger, Question, Neutral and Lombard, are used. Lombard effect speech is obtained by having speakers listen to 85 dB SPL pink noise through headphones while speaking (i.e., recordings are noise-free). A common vocabulary set of 35 aircraft communication words make up over 95% of the database. Twelve tokens of each word in the vocabulary are spoken by nine native American speakers for the Neutral conditions, and two tokens for each style condition. This database has been employed mostly in the study of how speech production and recognition varies when speaking under stressed speech conditions [94].

2.1 Speech Features for Speaker Recognition

This Section presents a brief review of features used for speaker recognition. A survey of automatic speaker recognition techniques is presented by Atal [4]. It has been shown that the cepstral coefficients are the best features for a speaker recognition system. Reynolds [16] has experimented with telephone-speech database for evaluation of performance of different features such as mel-frequency,

TH-1853_02610201

linear-frequency filter bank cepstral coefficients, LP cepstral coefficients and perceptual linear prediction (PLP) cepstral coefficients. Best results are reported for higher order linear prediction (LP) features and band-limited filter bank features. Campbell [6] has provided a comparison of different features such as reflection coefficients, log-area-ratios, arcsin reflection coefficients, line spectrum pair (LSP) frequencies and mel-frequency cepstral coefficients (MFCC). Evaluation of relative performance of different features such as LP cepstrum, autoregressive moving average (ARMA) cepstrum, log-area-ratios and reflection coefficients is carried out under different noise conditions [8]. LP derived cepstral coefficients have been used for a speaker-specific mapping for text-independent speaker recognition [19]. All these results show the wide acceptance of cepstral coefficients for speaker recognition applications. Integration of different feature sets for speaker recognition has been proposed by Chagnolleau et al [43]. This approach, named as time-frequency principal components (TFPC), seeks to find out the time dependent spectral vectors from the speech signal.

2.1.1 Linear prediction model based features

Linear prediction (LP) based features have wide applications in the area of speech science [2]. In linear prediction, the current speech sample is estimated as a weighted average of 'p' past speech samples. LP model is given as

$$\hat{x}[n] = \sum_{h=1}^{p} a_h x[n-h] \tag{2.1.1}$$

$\hat{x}[n]$ is the n^{th} predicated speech sample, a_h represents the h^{th} prediction coefficient and 'p' is the order of the linear predictor. LP coefficients are used to identify the speaker from his / her voice [6]. In addition to LP coefficients, four LP derived features, linear prediction cepstral coefficients (LPCC), arcsin reflection coefficients (ARC), log-area-ratios (LAR), and reflection coefficients (RC) are extensively used for speaker recognition applications. Two important algorithms to compute LP coefficients are Levinson algorithm and Burg algorithm. In this work, Levinson algorithm has been used to estimate the LP coefficients.

Linear Prediction Cepstral Coefficients (LPCC)

The linear prediction cepstral coefficients (LPCC) are the coefficients of the Fourier transform representation of the log magnitude spectrum. Atal [4] has shown a 70.3% speaker identification result with a database of 10 speakers. With the cepstral coefficients, a speaker verification error value of 2% is reported by Campbell [6] for 0.5 sec duration speech utterances. Inal et al [47] have shown a speaker

TH-1853_02610201

identification result of 91% with cepstral coefficients for a speaker database consisting of 7 males and 3 females. Cepstral study suggests that the lower order coefficients are more effective for automatic speaker recognition than the higher order coefficients [106]. The cepstral coefficients, which can be derived directly from linear prediction (LP) coefficients, are shown to be a robust and reliable feature set for speech analysis [2]. Mostly, these cepstral coefficients have superior quantization properties than impulse response and autocorrelation coefficients. Some form of preprocessing would limit the spectral dynamic range of the signal due to which the quantization properties show better results [113]. A linear prediction model effectively models the spectrum of the signal. The LP coefficients are further processed through the following algorithm [1] to derive cepstral coefficients. The conversion algorithm is given as

$$c_0 = \ln \sigma^2 \tag{2.1.2}$$

$$c_m = a_m + \sum_{k=1}^{m-1} (\frac{k}{m}) c_k a_{m-k}, \quad 1 \leq m \leq p \tag{2.1.3}$$

where σ^2 is the gain term in the LP model. Ten cepstral coefficients are estimated from each speech frame of 20 msec having an overlap of 10 msec in this work. This set of LPCCs ($LPCC_1$ to $LPCC_M$)is used as a cepstral feature vector for the speaker.

Reflection Coefficients (RC)

Reflection coefficients (RC) (k_m) can be derived as the lattice features of LP model [6, 114]. LP model based reflection coefficients are shown as equivalent to reflection coefficients of loss less acoustic tube model of the vocal tract [1, 107]. These features can be analyzed using a spectral sensitivity measure [115]. Reflection coefficients can be obtained as intermediate variables of Durbin's algorithm for solving LP model. A speaker identification result of 84% [47] is obtained with the reflection coefficients.

Log Area Ratio (LAR)

Viswanathan et al [115] proposed the log area ratio (LAR) as the nonlinear mapping of the reflection coefficients that approximately minimizes the average spectral sensitivity to quantization errors. The log area ratios are considered as features of the vocal tract and varies speaker to speaker. Log area ratio ($g_m = log[\frac{1+k_m}{1-k_m}], m = 1, 2, ..p$) is defined as the log of ratio of cross sectional areas of two

TH-1853_02610201

consecutive vocal tract segments. Furui has analyzed the speaker recognition with twelve log area ratio coefficients [58]. Initially, we have used ten coefficients for review analysis. Inal et al [47] have shown a speaker identification result of 80% with log area ratio coefficients.

Arcsin Reflection coefficients (ARC)

Linear quantization of the reflection coefficients is not satisfactory especially when some of them are close to unity in magnitude. Non-linear quantization is much more sensitive near 1 than near 0 [115]. To avoid the singularity of the LAR's at $k_m = 1$ while retaining approximately uniform spectral sensitivity [6], arcsin reflection coefficients (ARC) are used. The arcsin reflection coefficients are defined as $g_m^{'} = sin^{-1}k_m$, where k_m's are reflection coefficients.

2.1.2 Mel-Frequency cepstral coefficients (MFCC)

Mel Frequency Cepstral Coefficients (MFCC) [2,28] are the most widely used features for speech and speaker recognitions applications. MFCCs are estimated based on human perception of critical bandwidths. For calculation of MFCCs, filters spaced linearly at low frequencies and logarithmically at high frequencies are used to capture the phonetically important characteristics of speech. The mel-frequency scale is based on a linear frequency spacing below 1000 Hz and a logarithmic spacing above 1000 Hz. For calculation of mel-frequency spectrum, a bank of filters with each having a triangular bandpass frequency response (Fig.2.1) is used. The process of computing MFCCs is described as follows. After framing the speech signal, the next step in the processing is to window each individual frame so as to minimize the signal discontinuities at the beginning and end of each frame. In this work, a Hanning window is used. The next processing step is the computation of Fast Fourier Transform, which converts each frame of N samples from the time domain into the frequency domain. This frequency domain representation is referred as spectrum or periodogram. Psychophysical studies show that human perception of sound spectrum does not follow a linear scale. Thus for each tone with an actual frequency, f, a subjective pitch is measured on the mel-scale. As a reference point, the pitch of a 1 KHz tone, 40 dB above the perceptual hearing threshold, is defined as 1000 mels. The following formula is used to compute the mels for a given frequency f in Hz.

$$mel(f) = 2595 * \log(1 + \frac{f}{700}) \tag{2.1.4}$$

For calculation of mel-frequency spectrum, a bank of filters with triangular bandpass frequency

TH-1853_02610201

response is used. A constant mel-frequency interval is used as the bandwidth of the filters. The mel-spectrum or the output spectrum of $S(\omega)$ consists of the output power of these filters. The number of filters in the filter bank is chosen as 20 for this work. In the final step, MFCCs are estimated by evaluating the Fourier transform of the log mel-spectrum. In this work, from each speech frame of 20 msec with overlap of 10 msec, a set of MFCCs are computed. This set of MFCCs ($MFCC_1$ to $MFCC_M$) is used as a feature vector for the speaker.

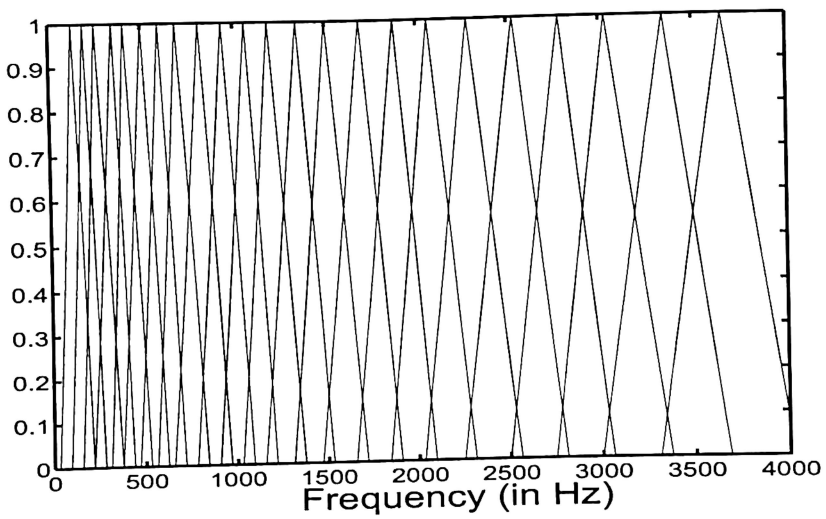


Figure 2.1: Mel-frequency filter bank

2.2 Statistical Feature Evaluation

One of the major objectives of this work is to analyze the performance of different speech features in characterization of speakers. Evaluation of relative performance of the features in characterization of the speakers will help this analysis. From literature, it is difficult to establish their relative performance as various authors have used different features and various classifiers for recognition of speakers. The error in recognition not only depends on features, but also on the nature of classification rules. It is therefore desirable to choose a measure which reflects the basic properties of the features to discriminate one speaker from the other and is not dependent on classification rules. Different speech features can be evaluated in order of their measured effectiveness to characterize the speakers. In this section, three statistical feature evaluation methods are discussed. These are probability density function (PDF) characteristics, analysis of variance or F-ratio test and Kolmogorov-Smirnov test.

TH-1853_02610201

2.2.1 Probability density function characteristics

Several utterances from the same speaker will generate a set of points in the feature space whose distribution can be described by a multivariate probability density function. Ideally, the feature vectors should have Gaussian distribution. With this assumption, the Gaussian mixture model (GMM) [45] has been proposed for speaker identification. The statistical models are usually HMMs or GMMs, which use multi-variant Gaussian probability density functions (PDFs) to represent the area in the feature space occupied by a particular speaker class. Each of these Gaussian PDFs are characterized by a mean vector and a covariance matrix which have been estimated from a set of training data. The feature space will be populated by many of these clusters which are determined by the number of speaker classes. For different speakers, the mean and variances of the feature vectors are different. As a result, the probability density characteristic is expected to be unique for each speaker. Speech feature will perform well, if these two factors have different values for various speakers. A feature set with a sharper probability density characteristic will produce better recognition accuracy [81].

The feature quality can analyzed by the probability density function characteristics. The probability density of first coefficient of six different features, arcsin reflection coefficients (ARC), reflection coefficients (RC), log area ratios (LAR), linear prediction cepstral coefficients (LPCC), LP coefficients (LP) and mel-frequency cepstral coefficients (MFCC), from speech data of four speakers are shown in Fig. 2.2. All the probability density function characteristics resemble Gaussian distribution. Four different characteristics are observed for the four speakers. The peaks of probability density characteristics lie in different positions. For the ARC feature three peaks are separable. This indicates that the classification can be done successfully for at least three speaker classes if this feature is used for recognition. This test is effective in finding the suitability of a feature coefficient for separating the speakers. This test is used in this work for analysis and evaluation of different speaker dependent speech features.

2.2.2 Analysis of variance

The pdf tests can show the effectiveness of a feature by graphical representation. It can be useful as a tool to evaluate the discriminating quality of a feature for few speakers. A quantitative evaluation method is needed to analyze the speech features' quality for speaker discrimination with large number of speakers. Features can be evaluated quantitatively using Fisher's discriminant ratio (F) [4]. Fisher's discriminant ratio or F-ratio test is defined as

TH-1853_02610201

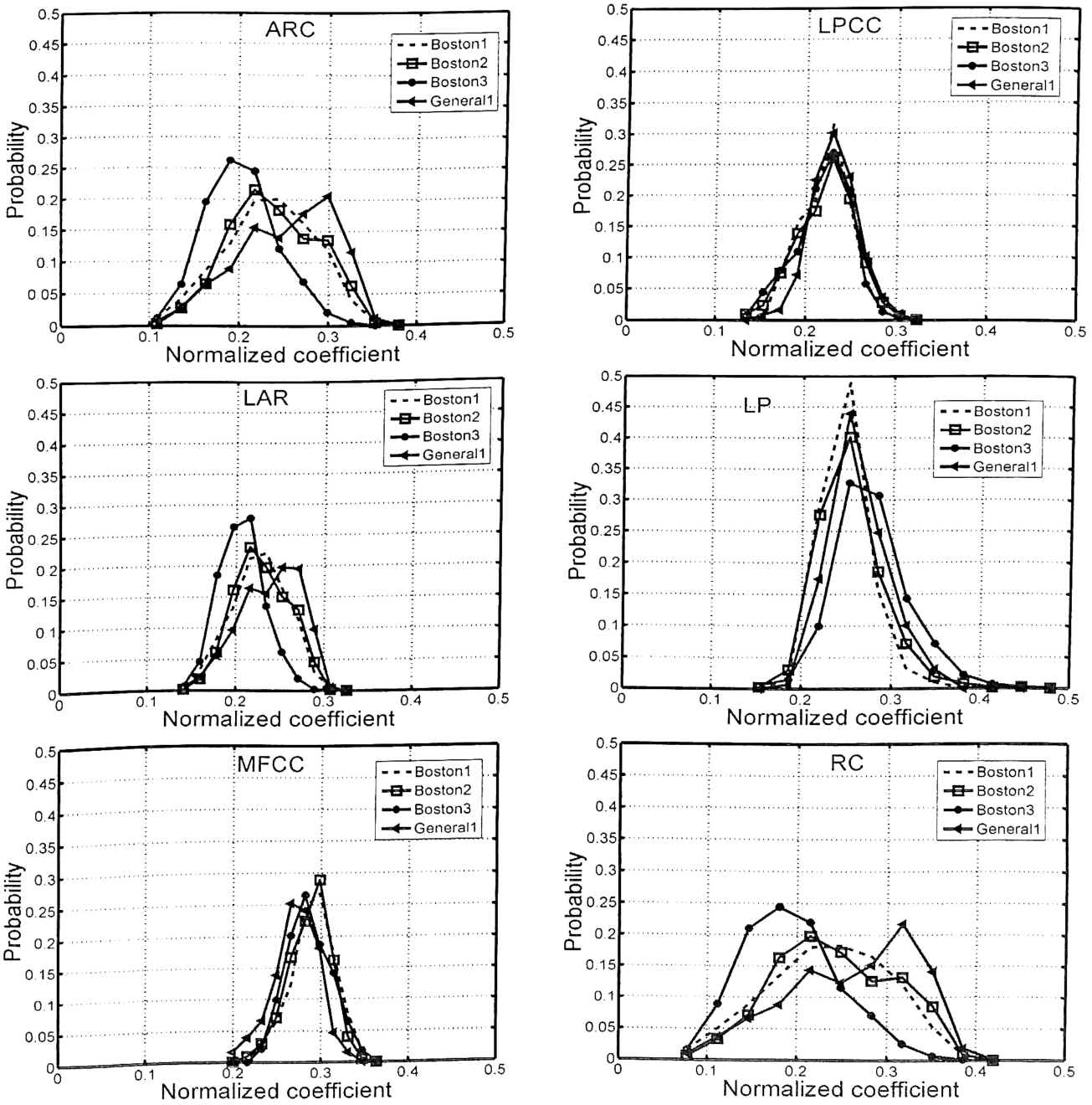


Figure 2.2: Probability density function characteristics

TH-1853_02610201

$$F = \frac{Variance\ of\ speaker\ means}{Average\ intra\ speaker\ variance} \tag{2.2.1}$$

In other words, the above ratio can be represented as,

$$F = \left\langle [\mu_i - \bar{\mu}]^2 \right\rangle_i / \left\langle [x_\beta^{(i)} - \mu_i]^2 \right\rangle_{\beta,i} \tag{2.2.2}$$

where $x_\beta^{(i)}$ is the feature value from the β^{th} utterance spoken by the i^{th} speaker, $<>_i$ indicates average over the speakers, $<>_\beta$ indicates average over all the utterances of the speakers, and

$$\mu_i = \left\langle x_\beta^{(i)} \right\rangle_\beta \tag{2.2.3}$$

is the estimated mean value of the feature for the i^{th} speaker, and

$$\bar{\mu} = < \mu_i >_i \tag{2.2.4}$$

is the estimated overall mean value of the feature averaged over all the speakers. Figure 2.3 shows F-ratio values of six features for nine speakers. Each graph shows F-ratio values of 12 coefficients of each feature. It is observed that the F-ratio values are different for the 12 coefficients of the same feature. Also, different features have different F-ratio values for the same coefficient. Higher F-ratio value will produce better recognition accuracy [4]. The F-ratio value may be high for features where one or two speakers are different from the rest, and when there exists correlation between coefficients of that particular feature set. Therefore, F-ratio values are most useful in eliminating poor features rather than choosing the best feature. The F-ratio value indicates the speaker discrimination ability of the speech feature. Features with higher F-ratio values do not always guarantee fewer recognition errors. In this work, the average of F-ratio values of all coefficients from a features is estimated for its evaluation. The relative performance of the speech features is evaluated by comparing the average F-ratio values of different features.

The major disadvantage of evaluating parameters individually is that interparameter correlations are not taken into account. For example, if the parameter is highly correlated with another parameter, the performance of the two together may not be better than either of the parameters alone. The individual evaluation of parameters can thus result in the selection of redundant parameters. It is advantageous to consider the parameters together as a vector. The concept of F-ratio can be extended to the multidimensional case. We define two covariance matrices: an interspeaker covariance matrix B and an intraspeaker covariance matrix w. These matrices are defined as

TH-1853_02610201

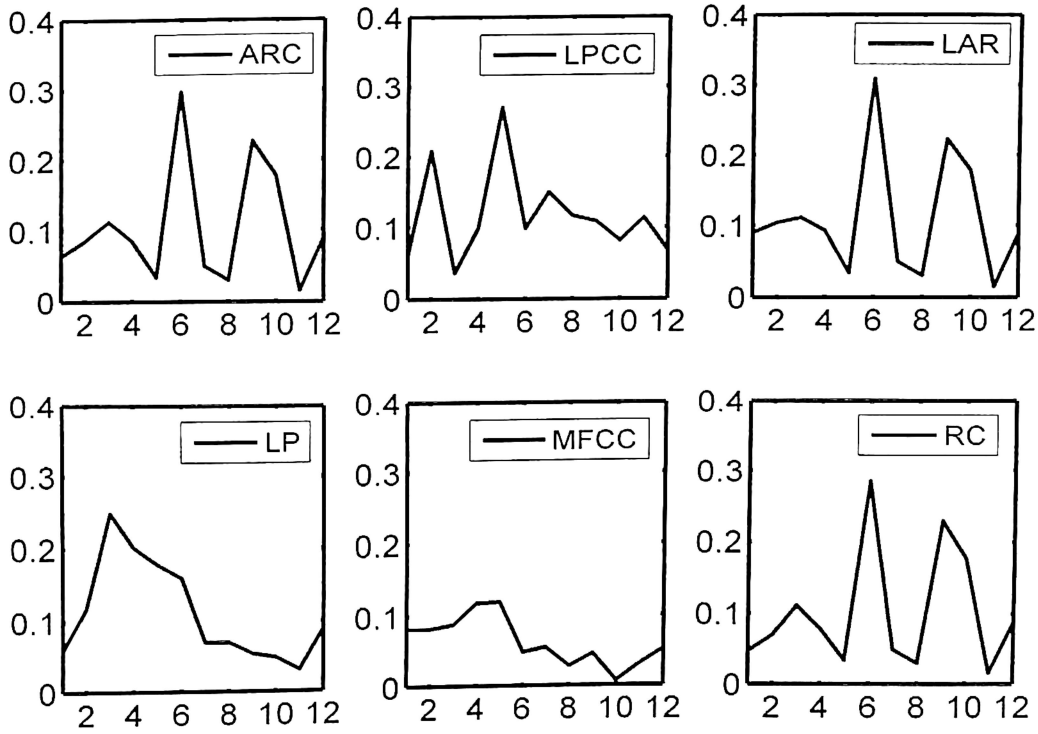


Figure 2.3: F-ratio values of six speech features evaluated for nine speakers

$$B = \left\langle [\mu_i - \bar{\mu}][\mu_i - \bar{\mu}]^t \right\rangle_i \tag{2.2.5}$$

$$w = \left\langle [x_\beta^{(i)} - \mu_i][x_\beta^{(i)} - \mu_i]^t \right\rangle_{\beta,i} \tag{2.2.6}$$

where $x_\beta^{(i)}$ is a vector representing the measurements from the β^{th} repetition of an utterance spoken by the i^{th} speaker, μ_i is the estimated mean vector of the feature for the i^{th} speaker and $\bar{\mu}$ is the estimated overall mean vector of the feature averaged over all the speakers. The divergence is given by [4]

$$Div = Tr\, w^{-1} B \tag{2.2.7}$$

where w^{-1} is the inverse of the matrix w and the superscript t indicates the transpose of a vector.

2.2.3 Kolmogorov-Smirnov (K-S) test

Kolmogorov-Smirnov [116] test is based on non-parametric statistical analysis. Nonparametric analysis allows user to analyze the data without assuming an underlying distribution and it is applicable to unbinned distributions which are functions of a single independent variable. In this work, each coefficient contour is calculated for two speakers. The maximum difference 'D' between these two contour

TH-1853_02610201

cumulative distribution functions (CDF) is estimated. K-S test is used to test whether two random samples are from the same (but unknown) distribution. Though there are an infinite number of normal distributions, each normal distribution is uniquely determined by its mean and standard deviation. Therefore, although speech features corresponding to every speaker came from the same family of distributions, they will have their unique normal distribution. In this work, K-S test is applied to determine the maximum difference between their empirical cumulative distribution functions of feature contours. The maximum difference between two CDF's can serve as the test statistic. The procedure for the test is,

1. First, the probability distribution function $f(x)$ is converted to a (CDF) $F(x)$. The value of $F(x)$ estimates the expected fraction of samples which are smaller or equal to x.

2. The maximum difference between two CDF's (D) serves as the test statistic, which is calculated as

$$D = \max_{-\infty<q<\infty} \left| F(x_q) - \hat{F}_n(x_q) \right| \tag{2.2.8}$$

and while comparing two distributions, the statistic would be

$$D = \max_{-\infty<q<\infty} |F_{n1}(x_q) - F_{n2}(x_q)| \tag{2.2.9}$$

An important result of K-S test is that it is invariant under reparameterization of x. In other words, the maximum distance D remains unchanged even if the horizontal axis values slide or stretch. A higher value of D indicates better recognition performance. Speech utterances of two speakers from the YOHO database are used for KS-test evaluation of six features. Figure 2.4 shows the maximum difference D between two speakers' CDFs for the first coefficient of six features. For each speaker, 96 speech utterances are used and each speech utterance is divided into sixty frames. For each frame, the average of the first coefficient over 96 utterances is chosen as the data sample for a CDF. Sixty average values of the first coefficient, which correspond to sixty frames, are used for determination of CDFs. These 60 values from the two speakers are used to find the maximum difference between the two CDFs. It is observed that the KS-test distance D is maximum for the LP coefficients. It is expected best speaker recognition performance will be observed if this coefficient is used for the recognition of the same speakers. To evaluate a feature, the average value of the maximum distances obtained from different coefficients is estimated and it is compared with the values obtained for other features.

TH-1853_02610201

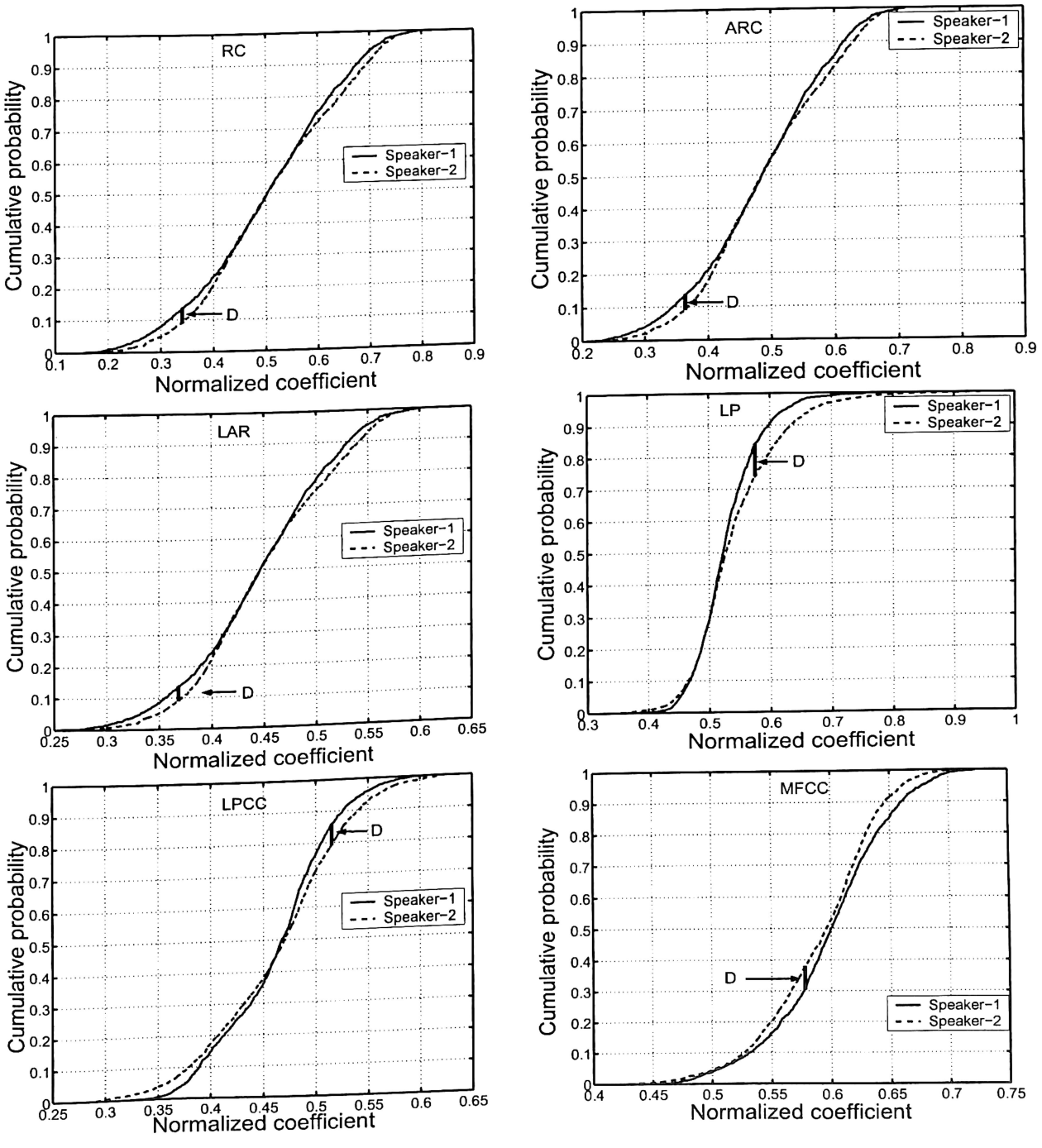


Figure 2.4: Maximum KS-test distance

TH-1853_02610201

2.3 Classifiers for Speaker Recognition

Researchers have used various classifiers for speaker recognition systems. Vector Quantization (VQ) [45], Hidden Markov Model (HMM) [46], Neural Networks [47] and Gaussian Mixture Model (GMM) [48] based classifiers are widely used for speaker recognition systems. In this work, non-parametric classifier, VQ, and parametric classifier, GMM, are used for evaluation of speaker recognition performance of different speech features.

2.3.1 Vector quantization (VQ)

This subsection presents the design of a vector-quantization (VQ) codebook [2] for speaker recognition. The speech signal is divided into frames and p number of features is extracted from the signal in each frame. Thus, for each frame, a set of p features, $x_t = [a_1, a_2, .., a_p]$, is generated and this represents one feature vector. From L number of frames of each recorded speech file, a set of input feature vectors given as $X = [x_1, x_2, ..x_T]$ is generated. In this way, possible feature vectors are extracted from the training speech data files for each speaker as training feature vectors. The codebook is designed based on the minimization of average distortion measure between the training vectors and the codebook Y ($Y = [y_1, y_2, ..y_P]$ and P is the size of the codebook). The average distortion measure is given as

$$\text{Average distortion measure} = \frac{1}{P}\sum_{v=1}^{P} min_{y_l \epsilon Y} d(x_v, y_l) \tag{2.3.1}$$

where $d(x_t, y_l)$ is the distance between the vectors x_t and y_l. Using the Generalized Lloyd algorithm, codebook is designed for each speaker with the training set of feature vectors from the speaker's data. These codebooks are the reference codebooks for recognition of speakers. A vector quantization classifier [5] shows 97% correct speaker recognition for a database of 10 speakers with the cepstral coefficients.

2.3.2 Gaussian mixture model

This subsection introduces the Gaussian mixture density, and factors which motivate the use of a Gaussian mixture model (GMM) as a modeling technique for speaker recognition, and describes a maximum likelihood procedure for Gaussian mixture parameter estimation. The discussion in this subsection will focus on GMM speaker model description and speaker classification for speech utterances. The GMM has several attributes which make it well suited for speaker recognition. First, the model

TH-1853_02610201

naturally arises from a probabilistic modeling of the underlying sounds which comprise a person's speech but does not depend on segmentation of the speech. By modeling the underlying sounds or acoustic classes, the model characterizes the short term variations of a person's voice and so is capable of high identification performance for short test utterances. GMM [16] basically models the feature vectors of the speaker as Gaussian densities. Then the likelihood ratio of speech segment being from a hypothesized speaker condition to not being from the hypothesized condition is calculated for speaker recognition. These GMMs are used for finding maximum a posteriori probability for a given speech segment.

For D-dimensional feature vectors, $\vec{x}$, the Gaussian mixture density is a weighted sum of M component densities and it is given by [86], $p(\vec{x}|\lambda) = \sum_{mix=1}^{M} w_{mix} p_{mix}(\vec{x})$. The above equation states that any density space can be represented as a weighted sum of Gaussian densities. Authors have taken weighted linear combination of M unimodal Gaussian densities. Each Gaussian density $p_{mix}(x)$ is parameterized by a mean vector $\vec{\mu_{mix}}$ which is a $p \times 1$ vector, and a $p \times p$ covariance matrix, Σ_{mix}. Since $p_{mix}(x)$ are Gaussian densities, they can be represented as

$$p_{mix}(\vec{x}) = \frac{1}{2\pi^{p/2}|\Sigma_{mix}|^{1/2}} \exp -\frac{1}{2}(\vec{x} - \vec{\mu_{mix}})'\Sigma_{mix}^{-1}(\vec{x} - \vec{\mu_{mix}}) \tag{2.3.2}$$

Furthermore, the mixture weights, w_{mix} , should satisfy the constraint $\sum_{mix=1}^{M} w_{mix} = 1$. The parameters of this density model are represented as $\lambda = \{w_{mix}, \vec{\mu_{mix}}, \Sigma_{mix}\}$ where $mix = 1, 2, \ldots, M$. For a given collection of training vectors, maximum likelihood model parameters are estimated using the iterative expectation-maximization (EM) algorithm. The EM algorithm iteratively refines the GMM parameters to monotonically increase the likelihood of the estimated model for the observed feature vectors, i.e., for iterations k and $k+1$, $p(X|\lambda^{k+1}) > p(X|\lambda^{k})$. Generally, five iterations are sufficient for parameter convergence. Usually, it is assumed that the feature vectors of X to be independent so the log-likelihood of a model λ for a sequence of feature vectors,$X = [x_1, x_2, ..x_T]$, is computed as $\log p(X|\lambda) = \frac{1}{T}\sum_{t=1}^{T} \log p(\vec{x}_t|\lambda)$. Note that the average log-likelihood value is used so as to normalize out duration effects from the log-likelihood value. Also, since the incorrect assumption of independence is underestimating the actual likelihood value with dependencies, scaling by T can be considered a rough compensation factor.

TH-1853_02610201

2.3.3 Speaker recognition system

The speaker identification system results in an enrolled speaker identity which is close to the testing utterance. Here, only one input testing utterance is enough. But the speaker verification system needs two inputs, which are speaker identity and testing utterance. Verification system has to decide genuine speaker or imposter speaker. Generally speaker identification is more complicated than verification. For classifier comparison, initially, we tested speaker identification. Speaker identification performance is analyzed using speaker identification rate. Speaker identification rate is calculated in terms of percentage of correctly identified testing utterances out of the total number of testing utterances.

Speaker verification system verifies the speaker information with a threshold. This threshold can be fixed using False Acceptance Rate (FAR) and the False Rejection Rate (FRR) [118]. FAR and FRR are functions of a threshold that can distinguish the two error rates. The performance of the verification system can be measured by plotting a Receiver Operating Characteristics (ROC) curve, which is a plot of FRR versus FAR. The point on the ROC defined by FAR=FRR is the Equal Error Rate point (EER). In speaker verification system less Equal Error Rate indicates better recognition. In speaker verification system, background speakers are selected to reduce the equal error rate. These background speakers' selection improves the speaker identification rate also. In literature, background speakers are used for speaker verification. These background speakers represent the population of expected imposters, which is in general application specific. Reynolds [16] has shown the speaker verification results with ten background speakers. Though cohort speaker selection is used for Vector Quantization based speaker recognition, far speaker set is not analyzed. So, the algorithm has been written for the Vector Quantization method with all constraints like close speaker set and far speaker set. Close speaker set used to increase the accuracy of the speaker recognition system. Here, the speakers who are similar to the genuine speaker are considered to get the best coverage. The background speakers selected as above are referred to as the maximally-spread close set. Similar to Reynold's far speaker set, we considered the maximally-spread far speaker set. When the Mahalanobis distance is used, the speaker recognition improves slightly with the far speaker set.

Speaker identification system with back ground speaker selection is described here. These back ground speakers are also used for speaker verification. The speaker identification system is the process of identifying the identity of unknown speaker's testing utterance. Here, the speaker set $S = \{1, 2, 3, ...N\}$ represents N number of speakers in a database. The matching scores of testing utterance with trained speakers provides the identity of the testing utterance. For the given speech feature vec-

TH-1853_02610201

tors $X = \{x_1, x_2, x_3, ..., x_T\}$ of testing utterance and codebook vectors $Y = [y_1, y_2, ..y_P]$ the average distortion can be computed between these two sets of vectors as follows

$$D(x_t, Y) = \min_{y_p \epsilon Y} \|x_t - y_p\|^2 \quad (2.3.3)$$

$$D_{avg}(X, Y) = \frac{1}{T} \sum_{t=1}^{T} D(x_t, Y) \quad (2.3.4)$$

The identity of the unknown speaker recognized by $i^* = \arg\min_i \{D_{avg}(X, Y_i) | i \epsilon S\}$ using this distortion measure. The performance of speaker identification system is improved by background speakers. For each speaker, background speakers like close speaker set and far speaker set are created and considered while identifying a speaker. The close speaker set and the far speaker set are used individually or in combination. Imposters will attempt to gain access only from similar sounding or at least same-sex speakers (dedicated imposters).

TRAINING UTTERANCE

TRAINING MODEL

SMSE $_{11}$	SMSE $_{12}$	SMSE $_{13}$	SMSE $_{1M}$
SMSE $_{21}$	SMSE $_{22}$		
SMSE $_{31}$			
SMSE $_{41}$			
○ ○ SMSE $_{N1}$			○ ○ SMSE $_{NM}$

Figure 2.5: Summation of Mean Square Error between training utterance and training Models(Codebook)

For the dedicated imposter case, selection of background speakers for the claimant speaker is done according to the distortion measure. The distortion measure between the codebook and training utterances is calculated. The training data of each speaker has been used for codebook preparation. The Sum of Mean Square Error (SMSE) between trained codebook and training data has been calculated. $SMSE_{NM}$ (Fig. 2.5) represents the sum of mean square error between the N^{th} speaker trained codebook and M^{th} speaker training utterances. If U_{12} indicates the 2^{nd} training utterance of 1^{st} speaker, the distortion measure between codebook and training utterance is $(D_{avg}(Y_1, U_{12}))$, where the sorted sum of mean square error provides the close speaker set and far speaker set. Here, I indicates the indices of

TH-1853_02610201

the sorted speaker set. From these indices, initial speakers are close speakers and last speakers are far speakers.

$$SMSE_{11} = D_{avg}(Y_1, U_{11}) + D_{avg}(Y_1, U_{12}) + ... + D_{avg}(Y_1, U_{1L}) \quad (2.3.5)$$

$$SMSE_{12} = D_{avg}(Y_1, U_{21}) + D_{avg}(Y_1, U_{22}) + ... + D_{avg}(Y_1, U_{2L}) \quad (2.3.6)$$

$$S_1 = [SMSE_{11}, SMSE_{12}, SMSE_{13}....SMSE_{1M}] \quad (2.3.7)$$

$$I = \arg sort(S_1) \quad (2.3.8)$$

Using the close speaker set and the far speaker set, speaker identification can be performed as follows

Table 2.1: Speaker identification results of YOHO database (106 male speakers, 12 coefficients, 256 size codebook) for different speech features

Speech features	MFCC	LAR	LPCC	ARC	RC	LP
Correct Identification	98.49	97.64	96.42	89.62	87.74	79.91

Table 2.2: Speaker identification results of neutral speech in SUSAS database (9 speakers, 12 coefficients, 32 size codebook) for different speech features

Speech features	MFCC	LAR	ARC	RC	LPCC	LP
Correct Identification	84.13	80.95	79.37	79.37	77.77	73.02

Table 2.3: Equal error rate (%) and false rejection rates at false acceptance rates of 0.1% and 0.01% for YOHO database

Database	M (10 css)	M (5 css and 5 fss)
YOHO (eer)	0.55%	0.62%
(FR @ FA = 0.1%)	3.00%	2.90%
(FR @ FA = 0.01%)	3.90%	4.12%

Tables 2.1-2.3 show speaker recognition results for YOHO and SUSAS databases. Speaker identification results are evaluated for the six speech features. The verification results in Table 2.3 is evaluated using MFCC features for the speech utterances in YOHO database. Speaker identification results using

TH-1853_02610201

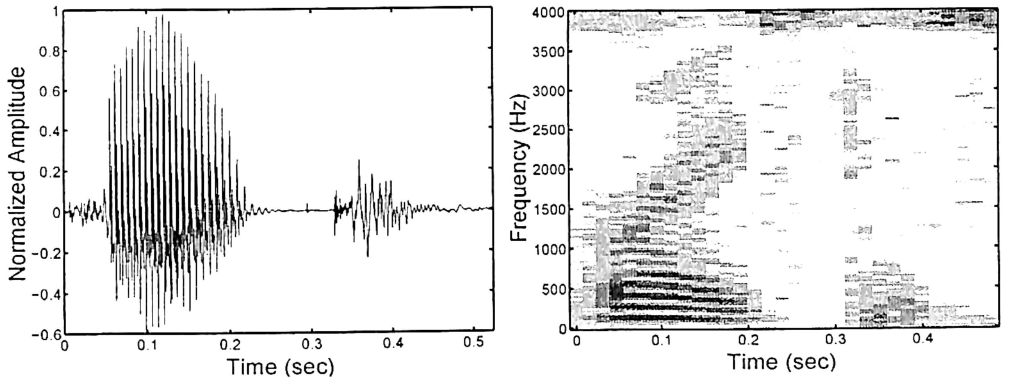
a 256-size VQ code book for the YOHO database is shown in Table 2.1. Speaker identification results using a 32-size code book for the SUSAS database is shown in Table 2.2. Reynolds [60] has shown a speaker identification rate of 99.7% for 106 male speakers from YOHO database and an equal error rate of 0.38. This equal error rate is achieved with 10 close speaker set. Ten true tests and 950 imposter tests are used for equal error rate calculation. For the same speakers, we obtained a VQ based speaker identification rate of 98.87% with the MFCC feature. Relative performance of the features show that MFCC feature has best performance and LP feature has the worst performance. For the YOHO database results, LAR, LPCC, ARC, RC and LP follows the MFCC in sequence of descending order of their performances. Similar relative performance is observed for the speech features in case of SUSAS database except for some small difference in the relative performance of ARC, RC and LPCC features.

2.4 Motivation for the Present Work

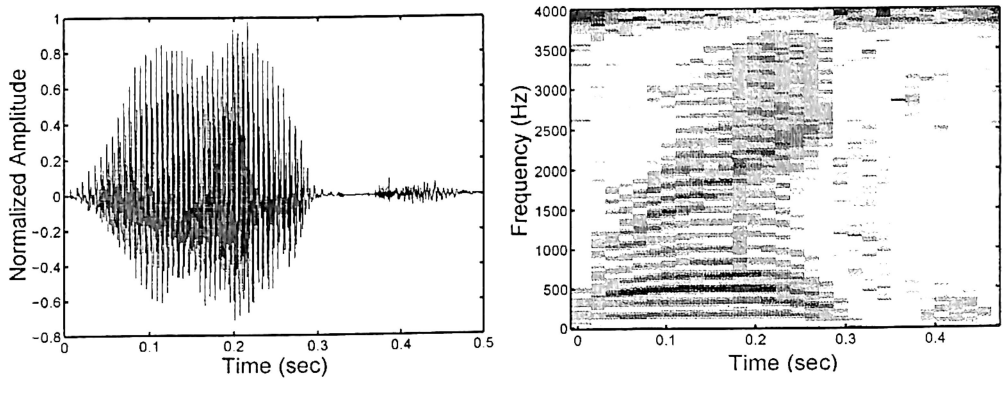
In this Section, we present the speaker recognition under stressed condition. One of the major causes of variability in a speaker's voice can be attributed to stressed condition. Stressed speech [84], [85], [86] is defined as the speech produced under any condition that causes the speaker to vary speech production from neutral condition. The causes of perceptually induced stress include emotion and environmental factors such as noise. Figure 2.6 shows two recorded stressed (emotional) speech signals and their spectrograms, one uttered with Neutral emotion and the other with Angry emotion. Both the signals are recorded from the same speaker. Some visible differences can be observed in terms of signal duration and signal energy. The speech uttered with Angry emotion has a duration more than that uttered with Neutral emotion. The total energy of the signal has a higher value in case of the speech signal uttered with Angry emotion. The spectrograms show that the relative energy in lower frequencies compared to higher frequencies is higher spectral magnitude in case of the speech uttered with Neutral emotion compared to the speech uttered with Angry emotion. The frequencies have shifted upwards or have higher values in the speech signal uttered with Angry emotion compared to the speech uttered with Neutral emotion. This shows that the characteristics of the speech signal vary under different emotions or stressed conditions.

Figure 2.7 shows histograms of 12 MFCC coefficients extracted from speech signals which are recorded under two emotions, Neutral and Angry, for the same speaker. Forty two speech signals are used for each emotion. For extraction of coefficients, 160 sample frame size with an overlapping of 80 samples is chosen for each speech signal. The coefficient values are normalized before histogram cal-

TH-1853_02610201



(a)



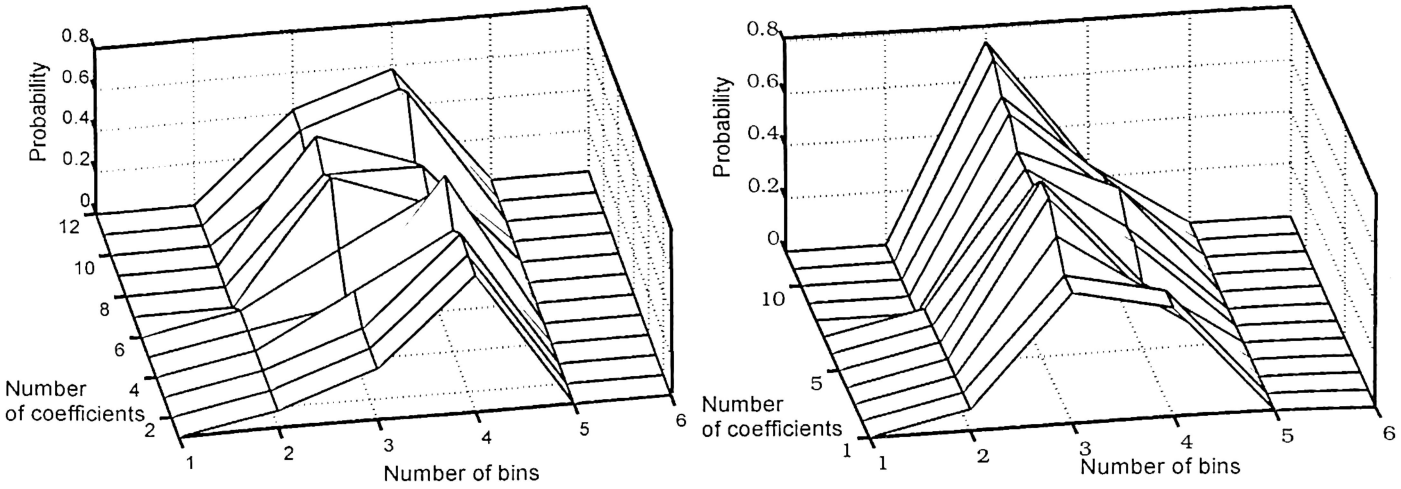
(b)

Figure 2.6: Speech signals (word: 'break') recorded from one speaker and their Spectrograms in two different emotions (a) Neutral (b) Angry

samples is chosen for each speech signal. The coefficient values are normalized before histogram calculation and six bins are used for each coefficient. It is observed that the histogram plots are different in two emotions for the same speaker. Coefficients, one, two, three, five, six, eleven and twelve, have their histogram peaks shifted from fourth bin to third bin between the emotions Neutral to Angry. Histogram plots prove that statistically these coefficients change their characteristics between the emotions Neutral and Angry. Due to changes in speech signal characteristics, the performance of speaker recognition system may degrade under stressed speech conditions if the codebook is trained with speech signals from a different stressed condition.

Normally, during the training sessions, speech signals are recorded in Neutral emotion condition. If the speech signal used for recognition is recorded in a different stressed condition, this speech signal characteristics may be different from that of the Neutral emotion condition. At the time of recognition, the stress condition will depend not only on the environment, which includes noise, but also on the person's state of mind. Table 2.4 shows speaker recognition results for speech signals recorded under

TH-1853_02610201



(a)

Figure 2.7: Histograms of MFCC coefficients from Speech signals (word: 'break') recorded from one speaker in two different emotions (a) Neutral (b) Angry

four different stressed conditions or emotions. Speech data corresponding to four emotions, Angry, Lombard, Neutral and Question from nine speakers of SUSAS database are used for the experiment. Twelve MFCC feature coefficients are extracted from each analysis frame of 160 speech samples with an overlapping of 80 speech samples. The number of MFCC feature vectors from each speech signal depends on the duration of the signal. A 256-size VQ codebook is designed for each speaker using 42 training utterances and the recognition results are evaluated using seven testing utterances from each speaker for each emotion. Four sets of codebooks, Angry, Lombard, Neutral and Question, corresponding to four emotions are implemented for each speaker. The testing utterances from each emotion are tested with the four sets of codebooks. Each row in Table 2.4 represents the codebook emotions and each column represents the testing emotions. Each value in the table shows the result for the speaker recognition value for testing utterances under the emotion indicated by the column and for the codebook used indicated by the rows. It is observed that the recognition results are different for the same codebook with test data from different emotions. Highest recognition result is obtained for Angry, Lombard, Neutral and Question trained data with Angry, Lombard, Neutral and Question test data respectively.We can observe it in terms of diagonal entries of the Table 2.4. Question codebook shows highest average recognition results. These results prove that the speaker recognition performance is different if the codebook data and the test data are from different stressed conditions.

The results obtained in Table 2.4 suggest that a detailed investigation is required for the performance of speaker recognition algorithms under stressed condition. During the last two decades, though some

TH-1853_02610201

Table 2.4: Speaker identification rate (%) for different emotions

Training (256	**Testing**				
size codebook)	Angry	Lombard	Neutral	Question	Average
Angry	71.42	41.26	41.26	36.5	47.61
Lombard	20.63	82.53	44.44	39.68	46.82
Neutral	30.15	49.2	76.19	60.31	53.96
Question	25.39	49.2	73.01	73.01	55.15

research publications in the area of recognition of stressed speech are published, there is little effort for analysis and recognition of speakers under stressed conditions. Studies on speech signals produced under different stressed and noisy conditions have helped improve the speech recognition algorithms [84, 99, 100]. These observations motivate the work taken up in this thesis.

Two major tasks in speaker recognition are feature extraction and classification. In this work, we investigate the effect of stressed condition on speech features. It is observed that the features which carry speaker dependent information are impacted due to stress condition. So, it will be useful to analyze the speaker dependent features under stressed conditions. It would be interesting to find out the relative performance of these features for speaker recognition under stressed conditions. Different speech features are tested for characterization of stressed conditions with various stressed speech databases. Though different classifiers show good recognition results for classification of different stressed conditions, they do not provide meaningful insight into the capabilities of various speech features in the characterization of different stressed conditions in speech signals. Both source and vocal tract are expected to be impacted during stressed conditions. Analysis of features from the point of view of source and vocal tract separately may help establish the fact whether the source or vocal tract is more effected by the stressed condition.

Speaker recognition systems use cepstral features such as mel-frequency cepstral coefficients (MFCC) [28], cepstral coefficients [106] and linear prediction (LP) cestral coefficients. MFCC features, which are the most effective among the three, are estimated from the spectral energies from different mel-frequency bands. In this way, the information associated with the spectral peaks (amplitudes and bandwidths) are not explicitly used. Other spectral domain features used for speaker recognition are LP coefficients and LP derived features, such as reflection coefficients, arcsin reflection coefficients and log-area ratios. The formants appear as spectral peaks in a preemphasized speech signal. Amplitudes, frequencies and bandwidths of the formants have significant speaker information. Above spectral do-

TH-1853_02610201

main features used for speaker recognition do not use these information explicitly. It is shown that performance of speaker recognition can be improved if different features are combined in a suitable manner [43]. A new set of spectral domain features, such as sinusoidal model based features, which can eliminate the above limitations will be useful towards improvement of performance of speaker recognition systems.

Various compensation techniques are used for improvement in speech recognition under stressed condition [48, 84, 94, 105, 108–110]. The spectral distortion induced by unusual speaking efforts could be compensated by simple linear transformation of the cepstrum [112]. The basic assumption of this technique was that spectral distortion induced by unusual speaking efforts could be compensated by simple linear transformation of the cepstrum. Although this assumption may be unduly optimistic, it was observed that the statistics of cepstral vectors did display some systematic modification in various speaking styles. The probability of spectral compensation in a stochastic model based recognizer was found to be effective in dealing with various articulation effects. Feature compensation techniques can be experimented for enhancement of performance of speaker recognition systems. The investigations presented in this thesis are broadly divided into three categories. They are as follows.

- First, the performance of the speaker recognition systems under stressed conditions is evaluated. Speech signals recorded under four different stressed conditions or emotions are used for evaluation of the performance. Different speech features are used for their relative performance evaluation. Also, analysis of speech features with preemphasized speech is carried out to compare the relative effect of stress condition on source and vocal tract.

- Literature show that the sinusoidal model based speech coder is well established for 4 kHz bandwidth transmission lines [119]. There is little reported work on sinusoidal model based speaker information analysis. In this work, sinusoidal model is tested for speaker recognition. Sinusoidal model has three features. They are amplitude, frequency and phase. Assuming that the information in these three features may be different, these features may be combined for improvement in speaker recognition. Also, sinusoidal features preserve the formant information as peaks in the spectrum is used for estimation of amplitude, frequency and phase features.

- Finally, compensation techniques are proposed and evaluated for improvement of performance of speaker recognition systems. Compensators are aimed at reducing the stress effect. As a result the compensated features are close to the speech features for normal or neutral condition. Four

TH-1853_02610201

novel compensation techniques are proposed and evaluated for improvement of speaker recognition under stressed condition. The compensation techniques are speaker and stressed information based compensation (SSIC), compensation by removal of stressed vectors (CRSV), cepstral mean normalization (CMN) and combination of MFCC and sinusoidal amplitude (CMSA) features. Speech data from SUSAS database corresponding to four different emotions or stressed conditions, Angry, Lombard, Question and Neutral, are used for analysis of speaker recognition under stressed condition. All the four compensation techniques help improve the speaker identification rates.

TH-1853_02610201

Chapter 3

Evaluation of Speech Features for Speaker Recognition under Stressed Condition

The performance of a speaker recognition system depends on the effectiveness of speech features in capturing the speaker dependent information. Features reflect the characteristics of the speech production system. These characteristics exist both in the spectral envelope (vocal tract characteristics) and in the supra-segmental features (voice source characteristics) of speech signal [9]. Spectral domain features are widely used for speaker recognition. Features which are commonly used for speaker recognition include mel-frequency cepstral coefficients (MFCC) [6] and linear prediction (LP) derived spectral features. Linear prediction (LP) coefficients and LP derived features, such as linear prediction cepstral coefficients (LPCC), reflection coefficients (RC), arcsin reflection coefficients (ARC) and log-area ratios (LAR), are used for speaker recognition. Performance of a speaker recognition system not only depends on the speech features and the classifiers, but also on the environment. For the same speaker, in a noisy environment the speech features may vary from the speech features extracted from speech signal recorded in a silent environment. In a noisy environment, the speaker is stressed. This causes the speaker to vary his or her speech production from Neutral condition. Performance of speaker recognition system will vary if the training speech data and the testing speech data are collected in different stress conditions.

In the previous chapter, six features, MFCC, LP coefficients, LPCC, RC, ARC and LAR, are discussed from speaker recognition point of view. The recognition results in Section 2.3 are based on the Neutral speech codebook. Analysis of speech features under stressed conditions will provide useful information which may be helpful in improving the performance of speaker recognition systems. For the same speaker, the vocal tract and source information are expected to vary for different stressed

TH-1853_02610201

conditions. Figure 3.1 shows spectrograms of speech signals under Neutral and Angry emotion. For each emotion, spectrograms with pre-emphasized speech and non pre-emphasized speech are shown for comparison. It is observed that pre-emphasis enhances the high frequency energy. Pre-emphasis suppresses or eliminates the glottal source and radiation effect [1]. For a pre-emphasized speech signal, LP model captures the vocal tract information and the excitation information is produced as the LP residual signal. In this work, speech features from pre-emphasized speech are evaluated for analysis of vocal tract information and speech features from non pre-emphasized speech are considered as representatives of both source and vocal tract information. In this chapter, we present the analysis and evaluation of six speech features, MFCC, LP coefficients, LPCC, RC, ARC and LAR, under stressed condition. Two classifiers, vector quantization (VQ) classifier and Gaussian Mmixture model (GMM) are used to evaluate the recognition results under Neutral and stressed conditions. Speech data of four stressed conditions, Neutral, Angry, Lombard and Question, are used for analysis. In Section 3.1, we present the statistical evaluation of the features. Probability density function (pdf) characteristics, F-ratio test and Kolmogorov-Smirnov (K-S) test are used for evaluation of speaker discrimination capability of the features under stressed conditions. Section 3.2 presents the analysis of the features based on vector quantization (VQ) and Gaussian mixture model (GMM) based classifiers. To assess the relative effect of stressed condition on vocal tract and source, the features from speech signals with pre-emphasis and without pre-emphasis are extracted. The relative ranking of the features based on their performance in speaker recognition under stressed condition is evaluated in Section 3.3. This analysis help select the best feature set for speaker recognition under stressed condition.

3.1 Statistical Evaluation of Speech Features

One of the major objectives of this work is to analyze the relative performance of different speech features in characterization of speakers. The error in recognition not only depends on features, but also on the nature of classification rules. It is therefore desirable to choose a measure which reflects the basic properties of the features to discriminate one speaker from the other and is not dependent on classification rules. Different speech features can be evaluated in order of their measured effectiveness to characterize the speakers. In this Section, three statistical evaluation methods are used for feature evaluation. These are probability density function (PDF) characteristics, analysis of variance (F-ratio) and a non-parametric test called Kolmogorov-Smirnov test.

TH-1853_02610201

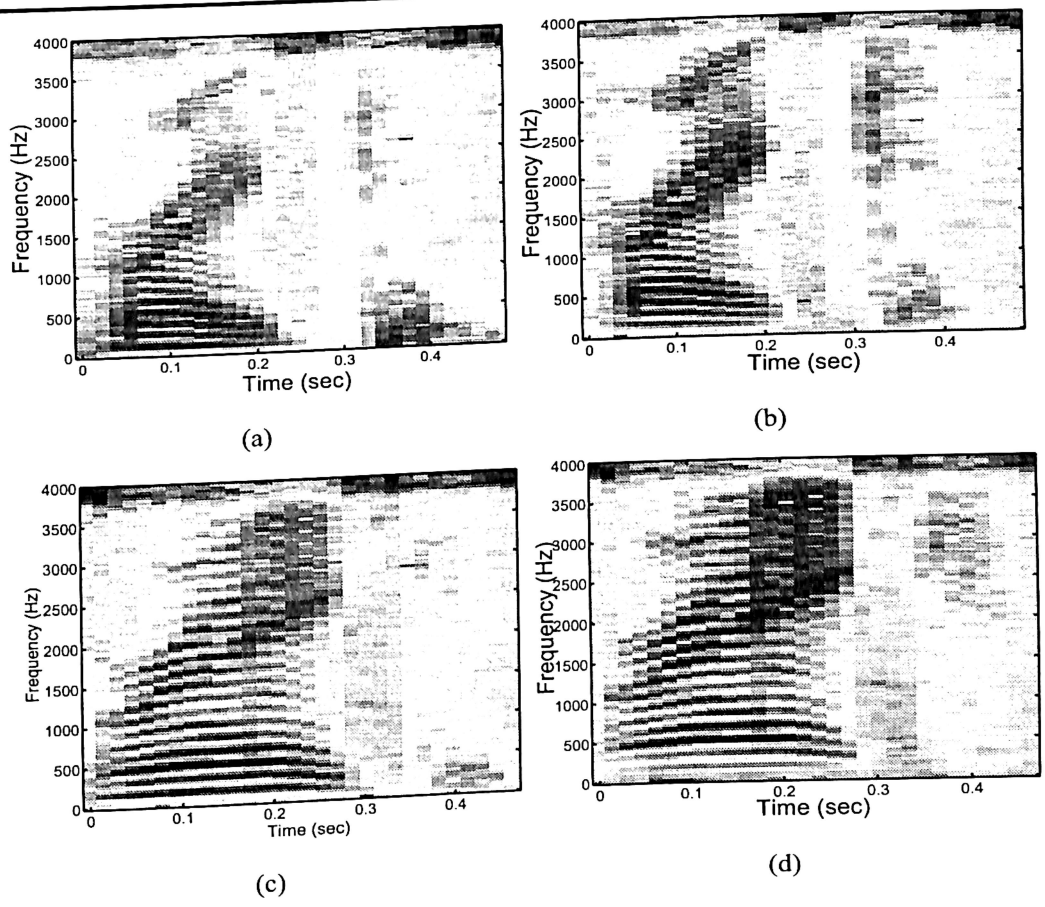


Figure 3.1: Spectrograms of speech signals (word: 'break') recorded from one speaker in two different emotions (a) Neutral (Non Pre-emphasized) (b) Neutral (Pre-emphasized) (c) Angry (Non Pre-emphasized) (d) Angry (Pre-emphasized)

3.1.1 Probability density function characteristics

Data distribution of a class close to the normal distribution leads to better classification [4]. probability density function (PDF) of a coefficient can be estimated for different speakers for evaluation of effectiveness of the coefficient. A coefficient will perform well if the mean values of the coefficients are different for various speakers. In this subsection, speech data of four speakers from SUSAS database are used for evaluation of PDF characteristics. A total number of 42 Neutral speech utterances per speaker is used for this purpose. Twelve coefficients per frame are estimated for each speech feature. To estimate the probability density function, the minimum and the maximum value of a coefficient are determined and then, the difference between them is equally divided into desired number of bins. In this work, ten bins are considered for evaluation of probability density functions. Probability density function of a coefficient is evaluated by estimating the histogram values of the coefficient.

Figure 3.2 shows that the probability density function characteristics of twelve linear prediction

TH-1853_02610201

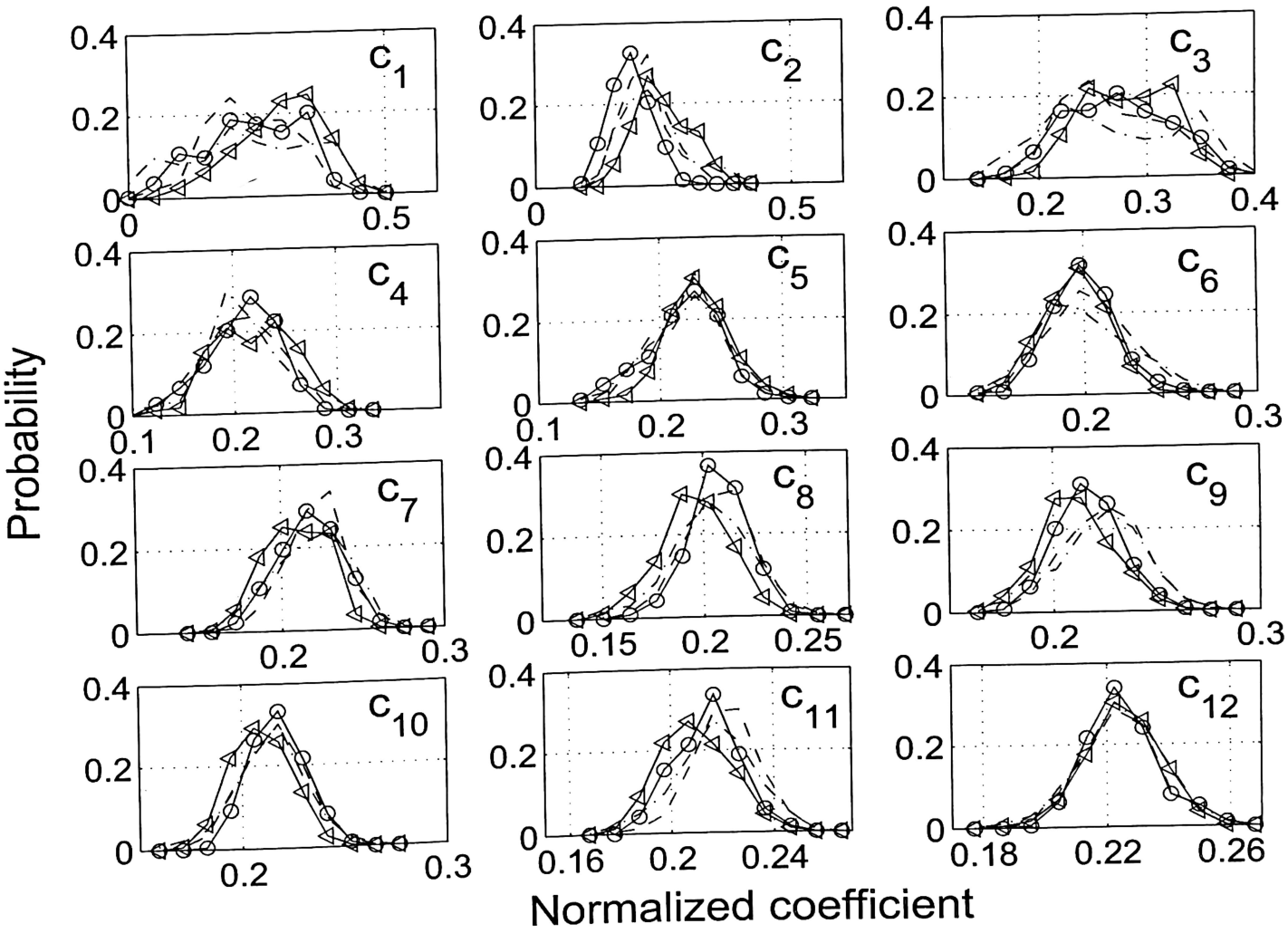


Figure 3.2: Probability density function of 12 LP cepstral coefficients (LPCC) for four speakers

cepstral coefficients (LPCC) for four speakers. It is observed that all the PDFs approximately resemble normally (Gaussian) distributed. In general, the PDF characteristics are different for four speakers. The peaks of the PDF curves are not at the same place which proves that the mean values are not same for the four speakers. The peaks look almost aligned for the coefficients C_6, and C_{12}. This suggests that these coefficients will be relatively less effective in discriminating the speakers compared to the other coefficients. The PDF characteristics suggest that the LPCC feature is capable of capturing the differences between speakers.

Speech feature variability due to stressed conditions can be evaluated by using the PDF characteristics estimated from speech signals under different stressed conditions or emotions. Figure 3.3 shows PDF characteristics of twelve linear prediction cepstral coefficients (LPCC) extracted from the speech data corresponding to four stressed conditions or emotions. The speech data are taken from the same

TH-1853_02610201

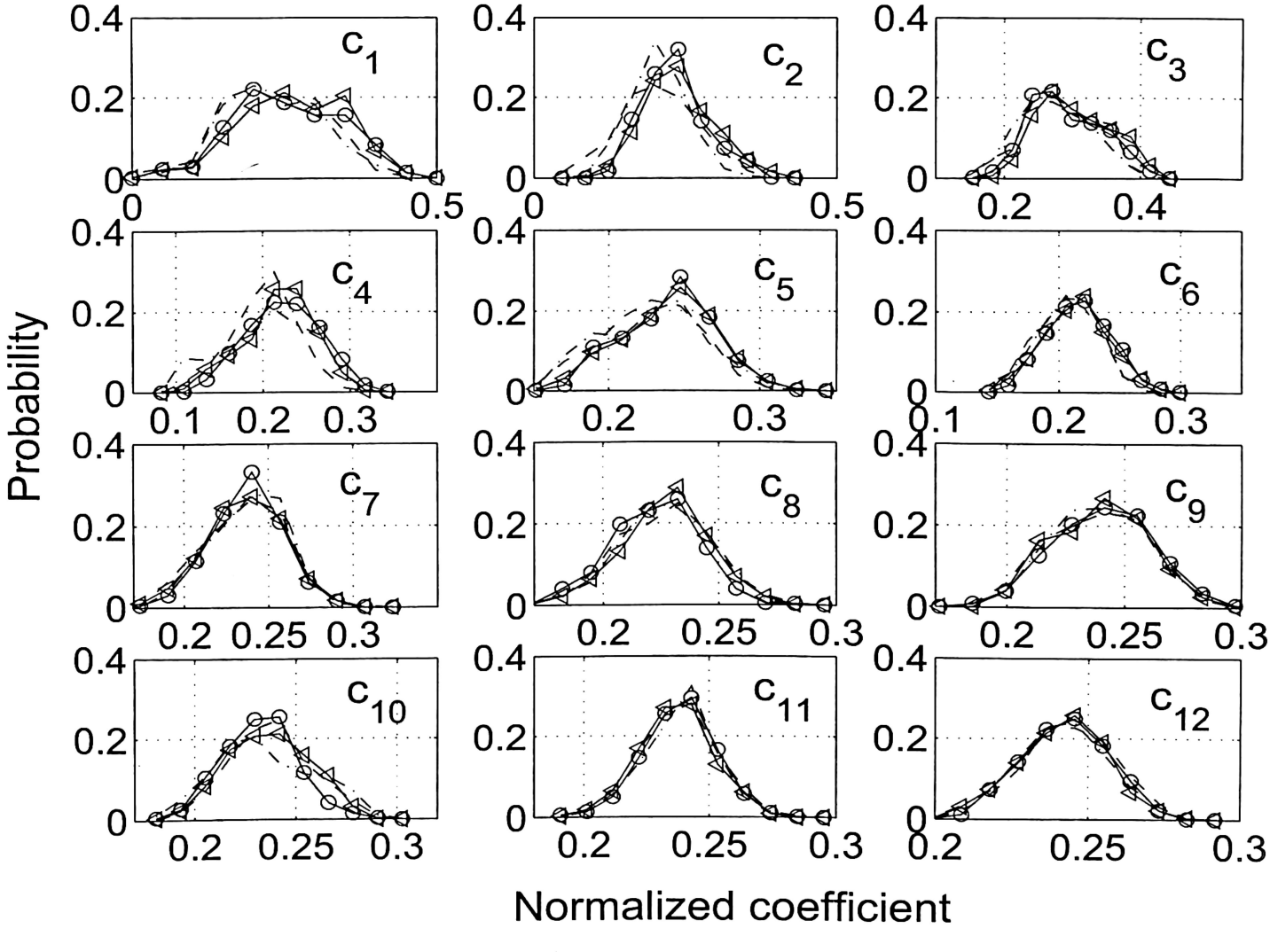


Figure 3.3: Probability density function of 12 LPCCs of four emotions data of a single speaker

speaker. It is observed that C_2 and C_4 coefficients have non-overlapping PDF characteristics. This suggests that these coefficients carry better emotion information compared to other coefficients for this speaker. The PDF characteristics for the C_{12} coefficient are almost aligned in both the figures, 3.2 and 3.3. This type of observation suggests that the C_{12} coefficient is neither suitable for speaker discrimination nor for classification of emotions. In general, for most of the coefficients, the PDF characteristics observed in Fig. 3.2 are different from the PDF characteristics observed in Fig. 3.3. This proves that the speaker information and the emotion information are captured with different degrees or in different manner by the coefficients. Similar types of results are observed from the PDF characteristics of other five speech features.

Above results show that the speaker information and the emotion information may be separated or one type of information may be enhanced compared to the other type of information. The results

TH-1853_02610201

shown with the probability density characteristics are interpreted visually. So the conclusions are based on subjective evaluation of PDF characteristics. These results are very useful for initial investigation on the properties of a feature. It may be difficult to evaluate the relative performance of different features by this method.

3.1.2 Analysis of variance

Both mean and variance values of the coefficients from a feature are important from the point of view of the feature's discriminating capability in a particular application. Analysis of variance by visual method is difficult. Quantitative evaluation of the performances is necessary to establish the relative importance of the features for speaker recognition. A feature set with higher F-ratio value will produce better recognition accuracy [4]. Each coefficient is separately evaluated for speaker recognition by F-ratio test. The speaker information distribution in feature space is evaluated using statistical terms as speaker mean and variance. F-ratio test is applied to all six speech features to test both the intra and the inter-distributions of speakers in feature space. The feature evaluation for speaker recognition is carried out under the Neutral and the stressed conditions. F-ratio values from speech data of nine speakers in SUSAS database are calculated for normal or non pre-emphasized speech for analysis of vocal tract and source information. F-ratio values estimated from pre-emphasized speech are used for analysis of vocal tract information. Higher F-ratio values suggest better recognition results. These results show the relative effectiveness of different coefficients of a feature for speaker recognition. In general, effectiveness of coefficients is different for the features. To select the effective feature, the average of F-ratio value over all coefficients for each feature is estimated.

Approximately, vocal tract information is obtained from pre-emphasized speech signal. Table 3.1 shows average F-ratio values for different features which are estimated from normal or non pre-emphasized speech signal and pre-emphasized speech signal. F-ratio values for nine speakers are estimated from speech signals under Neutral and Angry stressed condition. Comparing the F-ratio values in first and second columns, F-ratio values of LPCC and LP features in second column are more than those in first column. All other F-ratio values of pre-emphasized speech are higher than those from the normal or non pre-emphasized speech. This result suggests that all features, except LPCC and LP, may perform better for speaker recognition if they are taken from pre-emphasized speech. For the Angry speech data, comparing the F-ratio values in third and fourth columns, fourth column values are less than the third column values. According to these observations, the pre-emphasized speech would per-

TH-1853_02610201

Table 3.1: Average F-ratio values of different features

Average F-ratio of Neutral speech (pre-emphasized)	Average F-ratio of Neutral speech (non pre-emphasized)	Average F-ratio of Angry speech (pre-emphasized)	Average F-ratio of Angry speech (non pre-emphasized)
LAR(0.1154)	LPCC(0.1173)	LAR(0.0573)	LPCC(0.036)
ARC(0.1139)	LAR(0.1103)	ARC(0.0567)	LAR(0.0357)
RC(0.1127)	LP(0.1102)	RC(0.0561)	ARC(0.0341)
LP(0.0908)	ARC(0.1049)	LP(0.0539)	LP(0.0329)
LPCC(0.0783)	RC(0.1008)	LPCC(0.0402)	RC(0.0326)
MFCC(0.0627)	MFCC(0.0625)	MFCC(0.0343)	MFCC(0.0305)

form better for speaker recognition under stressed condition. It is also observed that the average F-ratio values for stressed speech signal are less compared to those from Neutral speech data. Hence, speaker recognition performance will be higher if the features are estimated from Neutral speech data compared to stressed speech data. F-ratio values of 12 coefficients of each feature are graphically shown in Fig 3.4, 3.5. Figure 3.4 shows that the F-ratio values of normal and pre-emphasized speech for nine speakers under Neutral condition. Solid line represents the non pre-emphasized speech and dashed line represents the pre-emphasized speech. The F-ratio values obtained from non pre-emphasized speech are different from those obtained from pre-emphasized speech. Figure 3.5 shows that the F-ratio values of normal and pre-emphasized speech for nine speakers under Angry stressed condition. Also for this stressed condition, the F-ratio values obtained from non pre-emphasized speech are different from those obtained from pre-emphasized speech. These results suggest that the speaker information for pre-emphasized speech is different from the non pre-emphasized speech. In majority of the cases (excluding LP and LPCC for neutral speech), features extracted from pre-emphasized speech contain better speaker discrimination information compared to features obtained from non pre-emphasized speech which is evident from the average F-ratio shown in Table 3.1.

F-ratio test helps identify the feature coefficients which are relatively less effective in classification. Analysis and comparison of F-ratio values between the speakers and between the stressed conditions

TH-1853_02610201

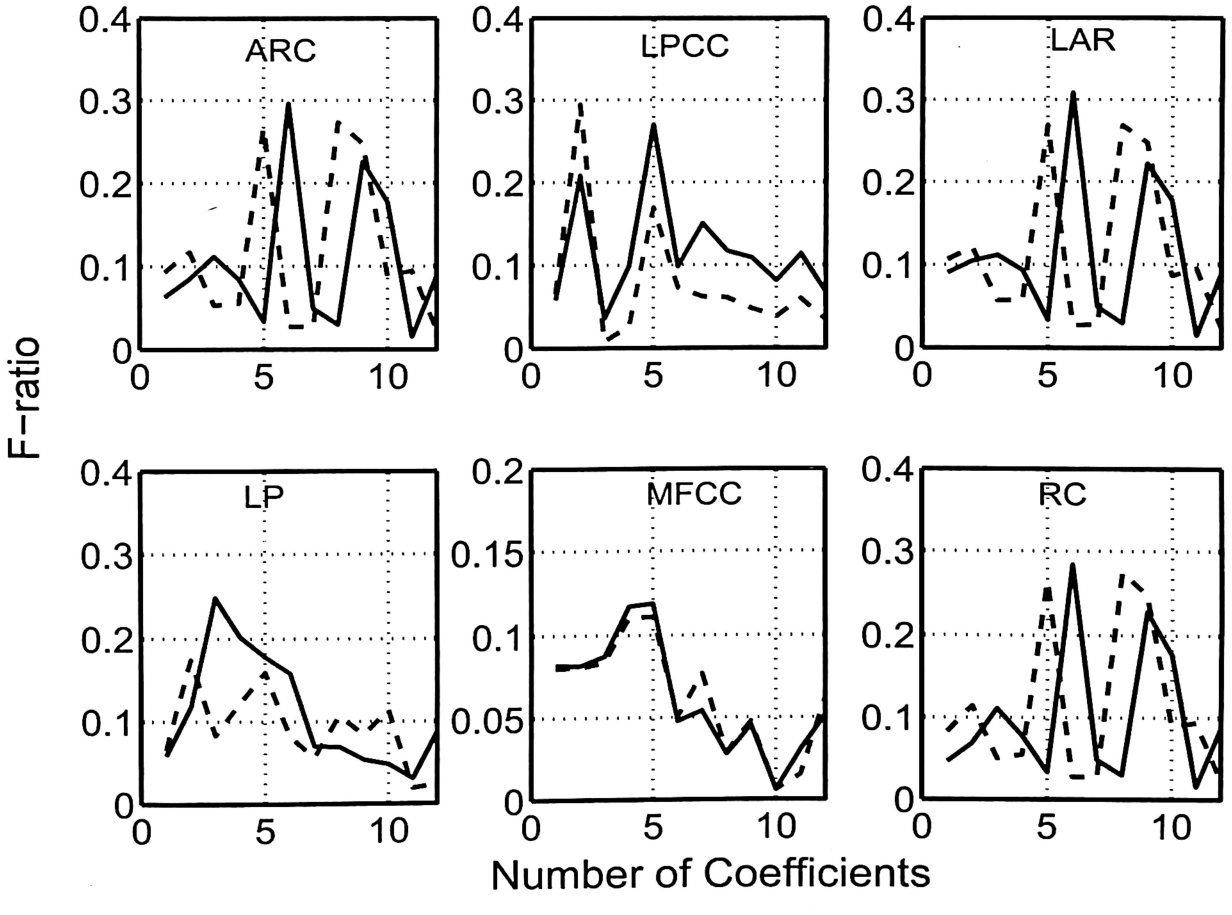


Figure 3.4: F-ratio values of six features of Neutral speech (Solid line - non pre-emphasized speech, Dashed line - pre-emphasized speech)

would give better idea about the information content in a feature coefficient. Higher F-ratio value among the F-ratio values evaluated between the speakers indicates that the feature coefficient would work well for speaker recognition. Similarly, higher F-ratio value among the F-ratio values evaluated between the stressed conditions indicates that the feature coefficient would work well for stress recognition. Figure 3.6 shows F-ratio values between speakers and between stressed conditions from the same speech data. It is observed that the F-ratio values between the speakers are different from the F-ratio values between the stressed conditions. An important observation is that for most of the features the relative F-ratio values between the coefficients show an inverse property. For example, F-ratio value of eighth coefficient of MFCC feature has a relatively higher value compared to other coefficients when evaluated between the stressed conditions. Same eighth coefficient of MFCC feature has a relatively lower value compared to other coefficients when evaluated between the speakers. This result suggest that this coefficient is a good carrier of information related to stressed conditions and it is a poor carrier

TH-1853_02610201

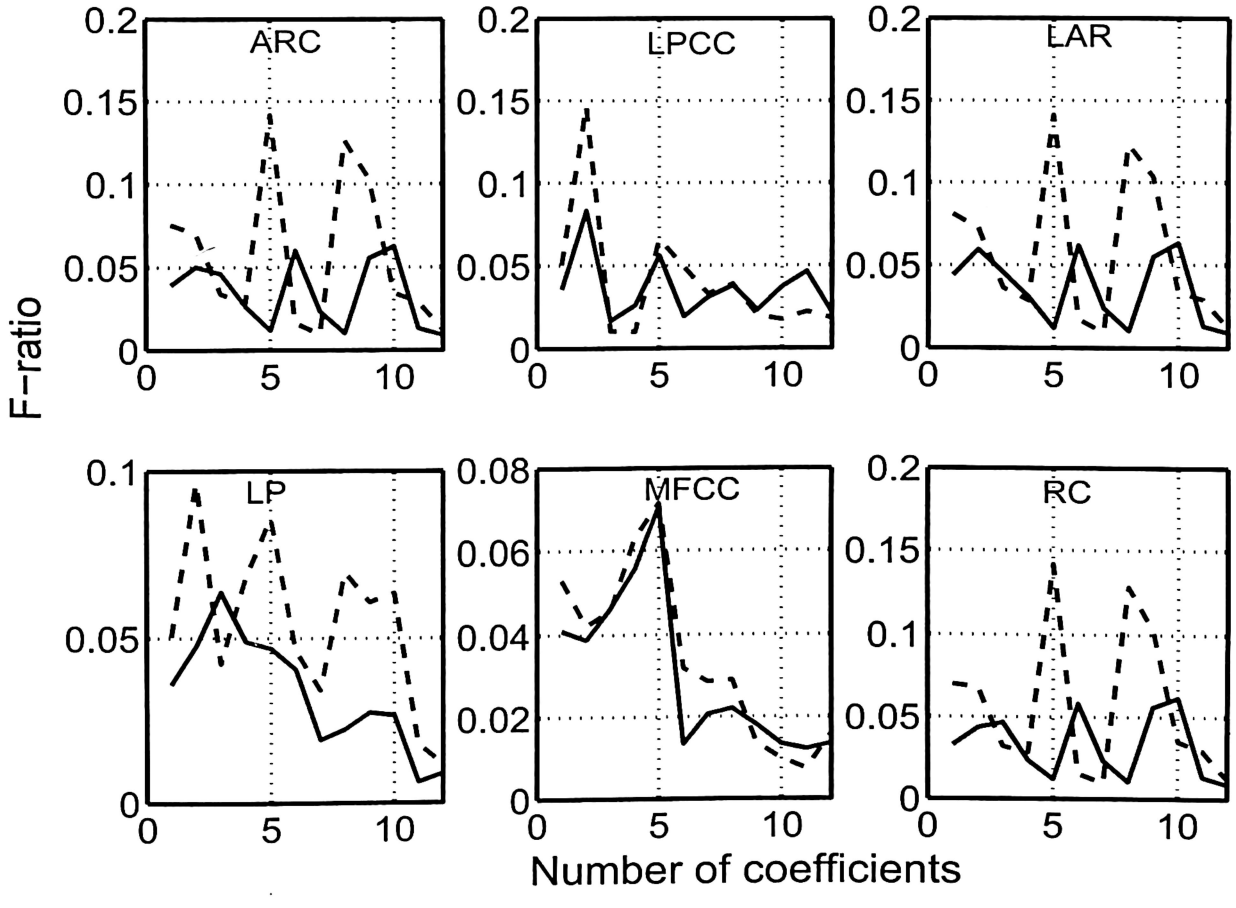


Figure 3.5: F-ratio values of six features of stressed speech (Solid line - non pre-emphasized speech, Dashed line - pre-emphasized speech)

of speaker information. By enhancing this coefficient value compared to other coefficients may result in improved recognition of stressed condition in a speech signal. Also, suppression of this coefficient may improve the speaker recognition performance because it will help eliminate the variations due to stress condition in a speaker's speech data. So, different coefficients can be weighed according to the applications for improvement in performance. This result proves that useful information for improvement in speaker recognition can be obtained by analysis of stressed conditions.

Comparison of F-ratio values and the speaker recognition results would give the idea about the feature ranking. F-ratio value tends to be high for features where one or two speakers are different from the rest, and when there exists correlation between coefficients of that particular feature set. Therefore F-ratio values may be useful in eliminating poor features rather than choosing the best. The major disadvantage of evaluating parameters individually is that inter parameter correlations are not taken into account. It is advantageous to consider the parameters together as a vector. The concept of F-

TH-1853_02610201

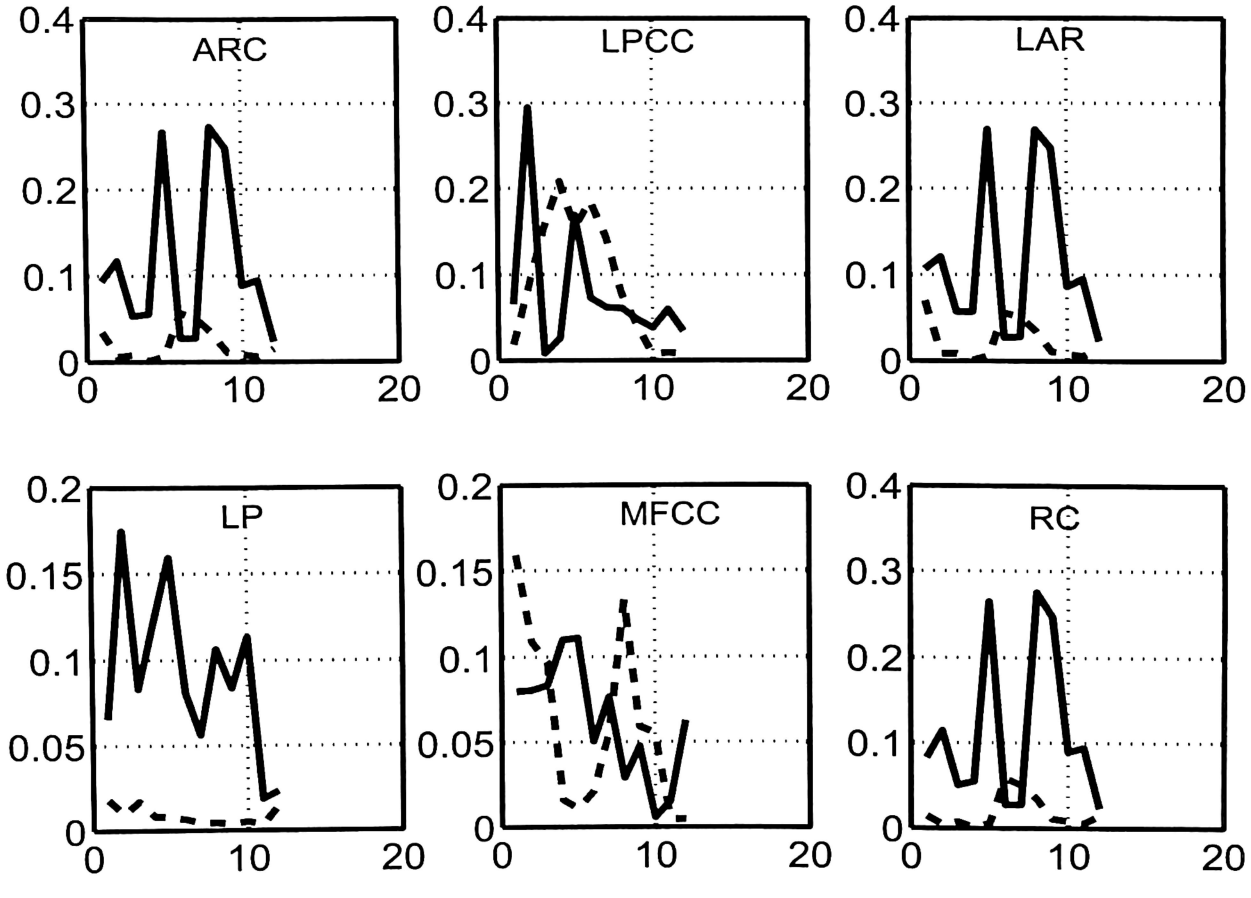


Figure 3.6: F-ratio values of six features (Solid line - speaker, Dashed line - stress)

ratio can be extended to the multi dimensional case. A suitable measure of effectiveness of a set of parameters for discriminating between different categories is the divergence which is a measure of uncertainty in the data. So, we have calculated the divergence of speech feature for different speakers. Divergence of non pre-emphasized and pre-emphasized speech are calculated for Neutral and Angry stressed speech. The features are ranked according to the divergence values and tabulated in Table. 3.2. The rankings of features for non pre-emphasized speech and pre-emphasized speech are not same. But the feature ranking in case of pre-emphasized Neutral speech (first column) and stressed speech (third column)are same. According to the divergence values from pre-emphasized Neutral speech (first column), the LAR feature is the best feature followed by ARC, RC, LP, MFCC and LPCC.

3.1.3 Kolmogorov-Smirnov Test

Kolmogorov-Smirnov (KS) [116] test is based on nonparametric statistical analysis. Nonparametric statistical analysis allows the user to analyze the data without assuming an underlying distribution. It

TH-1853_02610201

Table 3.2: Divergence values of six features

Divergence of Neutral speech (pre-emphasized)	Divergence of Neutral speech (non pre-emphasized)	Divergence of Angry speech (pre-emphasized)	Divergence of Angry speech (non pre-emphasized)
LAR(9.8098)	LP(18.5735)	LAR(4.8280)	LP(3.260)
ARC(9.6511)	LAR(10.4781)	ARC(4.7965)	LAR(3.080)
RC(9.5604)	ARC(9.4375)	RC(4.7672)	ARC(0.0341)
LP(8.6866)	RC(8.8645)	LP(4.6040)	LPCC(0.0329)
MFCC(8.0440)	LPCC(8.5619)	MFCC(3.6180)	RC(0.0326)
LPCC(7.5630)	MFCC(7.4868)	LPCC(3.4348)	MFCC(0.0305)

is applicable to unbinned distributions that are functions of a single independent variable. It is used to test whether two random samples are resulted from the same (but unknown) distribution. There may be an infinite number of normal distributions. But each normal distribution is uniquely determined by its mean μ and standard deviation σ. Therefore, although every speaker data came from the same family of distributions, they will have unique normal distribution. In this work, KS test is applied to calculate the probability of significance of the null hypothesis (that two speakers came from the same distribution), and the maximum difference between their empirical cumulative distribution functions. In Kolmogorov-Smirnov test, the maximum difference D between the two CDFs are calculated. The maximum difference D between the two CDFs indicate the overlapping of these two distributions. Higher value of D will suggest better recognition performance. In this work, the value of D is used for relative evaluation of the features. To analyze the speaker recognition performance, the maximum difference D between the possible two speaker pairs are calculated for each coefficient. Average of these differences of all these pairs is calculated. Average of all the coefficients of feature is calculated and tabulated in Table 3.3. Forty two utterances are considered here for both the speakers. It observed that the D values obtained from the pre-emphasized speech are higher compared to those obtained from the non pre-emphasized speech under the stressed condition. For the Neutral speech, expect for the LPCC, LP, and MFCC, other features have higher D values for pre-emphasized speech compared

TH-1853_02610201

to the non pre-emphasized speech. This result suggests that the pr-emphasized speech may produce better speaker discrimination compared to the normal speech. Values of D are higher for the features in Neutral speech compared to the stressed speech. Hence, for speaker recognition, features from Neutral speech will show better classification results compared to the features from stressed speech.

Table 3.3: Average of maximum distances between the two possible speakers CDFs

Average of CDFs of Neutral speech (pre-emphasized)	Average of CDFs of Neutral speech (non pre-emphasized)	Average of CDFs of Angry speech (pre-emphasized)	Average of CDFs of Angry speech (non pre-emphasized)
LAR(0.191)	LPCC(0.196)	LP(0.141)	LPCC(0.122)
ARC(0.191)	LP(0.189)	LAR(0.141)	MFCC(0.112)
RC(0.191)	LAR(0.189)	ARC(0.141)	LAR(0.109)
LP(0.188)	ARC(0.188)	RC(0.141)	ARC(0.109)
LPCC(0.160)	RC(0.188)	MFCC(0.118)	RC(0.109)
MFCC(0.153)	MFCC(0.154)	LPCC(0.116)	LP(0.105)

KS test is analyzed for normal speech and pre-emphasized speech under Neutral and stressed conditions. The we evaluate KS-test of speech features for speaker recognition under stressed condition. For this analysis, the maximum difference D has been calculated between the same speaker's different emotion distributions. KS-test is carried out between speaker's reference emotion to same speaker's other emotions. Average of these differences is calculated for each speaker and averaged for all nine speakers. Average of these maximum differences between the emotions within the speaker is calculated for comparison and tabulated in 3.4. A feature with a lower value of average maximum difference will indicate its better performance for speaker recognition. From Table 3.4, it is observed that for all the features the distance between the speaker's emotion is least when Neutral emotion is considered as the reference emotion. It can be concluded that for a feature a high average KS test distance between the speakers and a low average KS test distance between the speaker's different emotions indicate better performance for speaker recognition. The average value in the fifth column shows the feature variability under stressed condition. The LP coefficients show the least variability. MFCC and LPCC features

TH-1853_02610201

show relatively more variations within the speaker (intra speaker). These results suggest that MFCC and LPCC features may result in better emotion classification.

Table 3.4: Average of Maximum distances between CDFs of a speaker's reference stress condition and all other stress conditions

Speech Feature	Emotions				Average
	Angry	Lombard	Neutral	Question	
LP	0.0551	0.0325	0.0296	0.0331	0.0376
LAR	0.0573	0.0422	0.0351	0.0405	0.0438
ARC	0.0573	0.0422	0.0351	0.0405	0.0438
RC	0.0573	0.0422	0.0351	0.0405	0.0438
MFCC	0.0756	0.0452	0.0450	0.0450	0.0527
LPCC	0.0731	0.0565	0.0494	0.0559	0.0587

3.2 Feature Analysis using Classifiers

The purpose of analysis of speech features is to find out the reasons for degradation of performance of speaker recognition systems. Through statistical evaluation techniques, discussed in the previous Section, the relative differences between the features from different stressed speech signals are estimated. Evaluation of relative recognition performance of features will help choose the features which are robust to the variations in the speech signal due to different stressed conditions or emotions. In this Section, we evaluate the recognition performance of six speech features, MFCC, LP coefficients, LPCC, RC, ARC and LAR, for speech signals under Neutral and stressed conditions. Two classifiers, vector quantization (VQ) and Gaussian mixure model (GMM) are used for relative performance evaluation of the the features.

3.2.1 Vector quantization (VQ) classifier

In this subsection, vector quantization (VQ) classifier [2] is used for evaluation of relative speaker recognition performance of different speech features. Speaker identification and verification perfor-

TH-1853_02610201

mances of the features are evaluated with the vector quantization classifier. Sixteen size codewords are prepared for these comparison. Using the Euclidean distance measure, every vector is clustered around each codeword. The input vector belongs to the cluster of the codeword that yields the minimum distance between them. Following the above steps, a new set of codewords have been computed by estimating the average of each cluster. By repeating the above two steps until the iterations are exhausted, a final set of code words is calculated.

From SUSAS database, a total of 42 Neutral utterances per speaker are used for training and 7 stressed speech data files per emotion including Neutral are used for testing. The testing phase includes loading all the code books and then finding the minimum distant codebook. The results in terms of speaker identification rate are shown in Table 3.5. Each column shows the speaker identification rate for the testing speech utterances from a particular emotion. The speaker models are trained by Neutral speech utterances. Last column shows the average speaker identification rate of the feature and they are shown in descending order. In this experiment, 12 coefficients are considered for each feature vector. As the codebook is trained with Neutral speech utterances, the recognition results show maximum value in case of Neutral testing utterances for all the features. Question emotion testing utterances show higher recognition results compared to Angry and Lombard emotion testing utterances. ARC features show best average performance among all the features. Table 3.6 shows the results for the features extracted from normal or non pre-emphasized speech signals. Similar to the previous experiment, Neutral codebook is used for evaluation of recognition results. Neutral emotion test utterances show best result followed by Question stressed speech test utterances. Higher recognition results for Question test utterances may be due to the fact that the voiced speech components are emphasized under this stressed condition. Comparing the results in Tables 3.5 and 3.6, it is observed that all the features perform better if they are taken from the pre-emphasized signal. Pre-emphasized speech signal has vocal tract and excitation information. Radiation filter and glottal filter effects are approximately filtered out by pre-emphasizing the signal. The above results prove that isolating vocal tract and excitation improves the speaker recognition results not only under Neutral condition but also under other stressed conditions.

Figures 3.7, 3.8, 3.9, 3.10, 3.11 and 3.12 show speaker verification results in terms of genuine speaker scores and imposter speaker scores. Nine speakers' Neutral data is used for training and testing of above speaker verification results. A 16 size VQ Code book for is prepared for each speaker using 42 training utterances and seven utterances from each speaker are used for testing. For each feature, 63 (7×9) genuine scores are obtained corresponding to seven test utterances from each of nine speakers. Similarly, 504 ($7 \times 9 \times 8$) imposter scores are obtained from seven test utterances from each of nine

TH-1853_02610201

Table 3.5: Speaker identification rate using VQ classifier with Neutral emotion codebook (pre-emphasized speech)

Features	Speaker identification rate (%)				
	Emotion of testing utterances				Average
	Angry	Lombard	Neutral	Question	
ARC	31.74	44.44	79.36	63.49	54.75
MFCC	26.98	49.2	80.95	58.73	53.96
RC	33.33	41.26	82.53	58.73	53.96
LAR	22.22	46.03	79.36	55.55	50.79
LPCC	28.57	46.03	71.42	46.03	48.01
LP	31.74	41.26	58.73	42.85	43.64

speakers which are used for eight speaker code books. Genuine scores near value of one and imposter scores clustered near zero will indicate better performance of the features. Out of the six features, MFCC feature (Fig. 3.8) shows that the genuine speaker scores and the imposter speaker scores are well separated compared to the other features. The speaker verification results with pre-emphasized speech under Neutral and stressed conditions are shown as DET curves in Fig. 3.13 and Fig. 3.14 respectively. For Neutral speech, MFCC feature has the lowest equal error rate (EER) value and LP feature has the highest EER value. This shows that MFCC feature will perform best for speaker verification under Neutral condition. For the stressed speech, the EER values for all the features are more compared to the Neutral speech. The relative performance of all the features is same in both the stressed speech data and the Neutral speech data.

Table 3.7 shows speaker identification (SI) and speaker verification (SV) results for Neutral speech. The results show that the SI rate of non pre-emphasized speech is less than the pre-emphasized speech. In case of speaker verification, ARC and LAR features show higher EER value for pre-emphasized speech signals compared to non pre-emphasized speech signals. Table 3.8 shows SI and SV results for stressed speech signals with the reference speaker models trained with the Neutral speech. It is observed that MFCC feature shows best performance for SV results. In case of SI, all the features show

TH-1853_02610201

Table 3.6: Speaker identification rate using VQ classifier with Neutral emotion codebook (non pre-emphasized)

Features	Speaker identification rate (%)				
	Emotion of testing utterances				Average
	Angry	Lombard	Neutral	Question	
RC	42.85	42.85	80.95	47.61	53.57
ARC	38.09	36.5	80.95	50.79	51.58
LAR	22.22	39.68	80.95	60.31	50.79
MFCC	20.63	38.09	79.36	50.79	47.22
LPCC	22.22	39.68	76.19	47.61	46.43
LP	17.46	33.33	65.07	47.61	40.87

better performance with the pre-emphasized speech compared to the non pre-emphasized speech. In case of SV, expect for MFCC features, all other features show lower EER value for pre-emphasized speech. The speaker verification result is more prompted than speaker identification results because of decision made by a single threshold. According to the verification results, we can conclude that pre-emphasized speech performs well for speaker recognition under stressed condition.

Table 3.9 shows SI results for MFCC features with a Neutral trained code book. MFCC features are extracted from both pre-emphasized and non pre-emphasized speech. This experiment is performed to verify the SI performance when the number of MFCC features is increased. Speaker identification results are obtained with two different number of MFCC features. A 256 size code book has been prepared for each speaker. The results show that the pre-emphasized speech show better speaker identification rate for both the twelfth and twentieth number of MFCC features. The improvement is not significant when the number of MFCC features is increased from twelve to twenty.

Table 3.10 and Table 3.11 show speaker identification results for the six features with different stressed conditions. Table 3.10 shows results for features extracted from non pre-emphasized speech and Table 3.11 show results for features extracted from pre-emphasized speech. For each feature four 256 size code books are prepared using features from four emotions or stressed conditions. Then

TH-1853_02610201

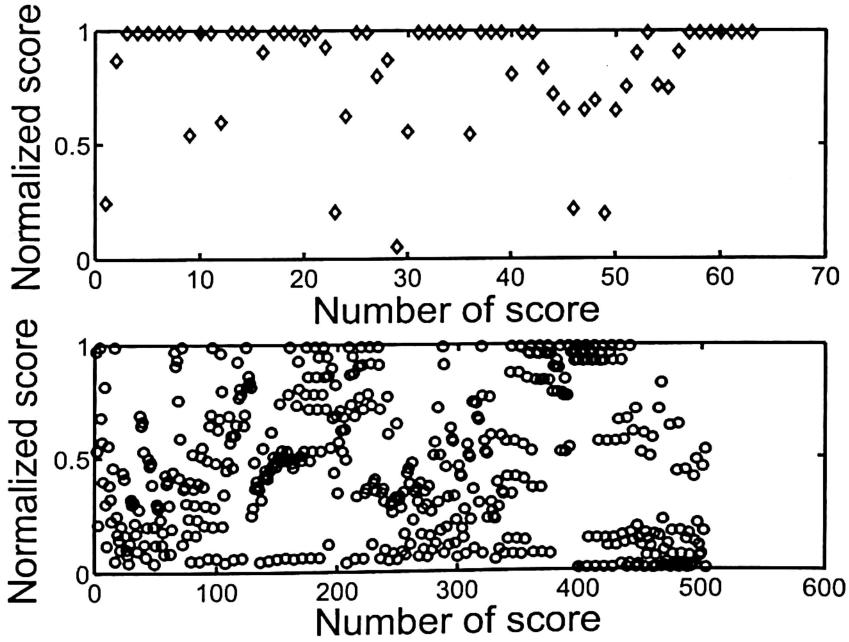


Figure 3.7: Genuine (diamond symbol) and Imposter (circular symbol) score of Log area ratios

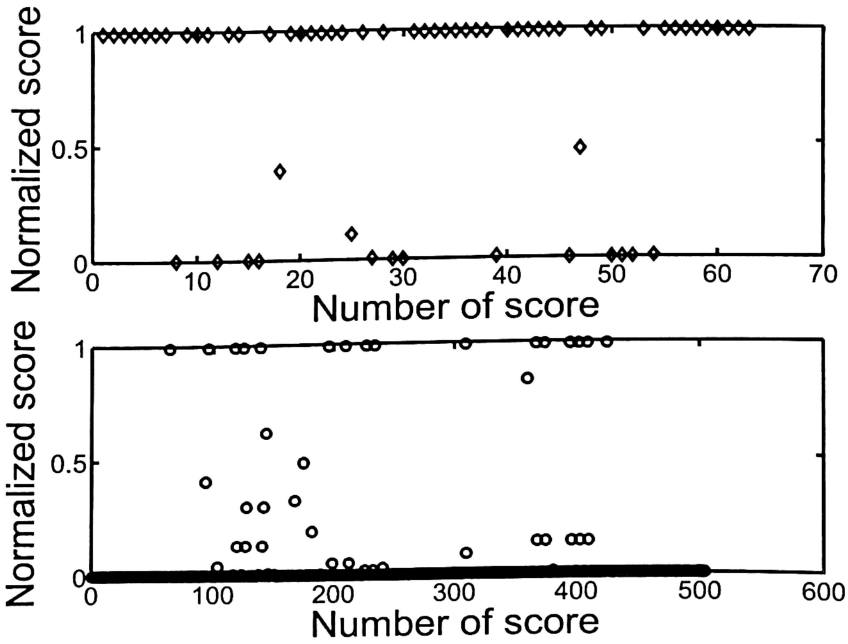


Figure 3.8: Genuine (diamond symbol) and Imposter (circular symbol) score of Mel-frequency cepstral coefficients

speaker identification results are evaluated using test features extracted from the four emotion or stressed speech data. In Table 3.10, it is observed that the Question emotion based trained models show maximum average recognition followed by Neutral, Lombard and Angry for all emotion testing utterances for ARC, LAR and MFCC features. In Table 3.11, it is observed that the Question emo-

TH-1853_02610201

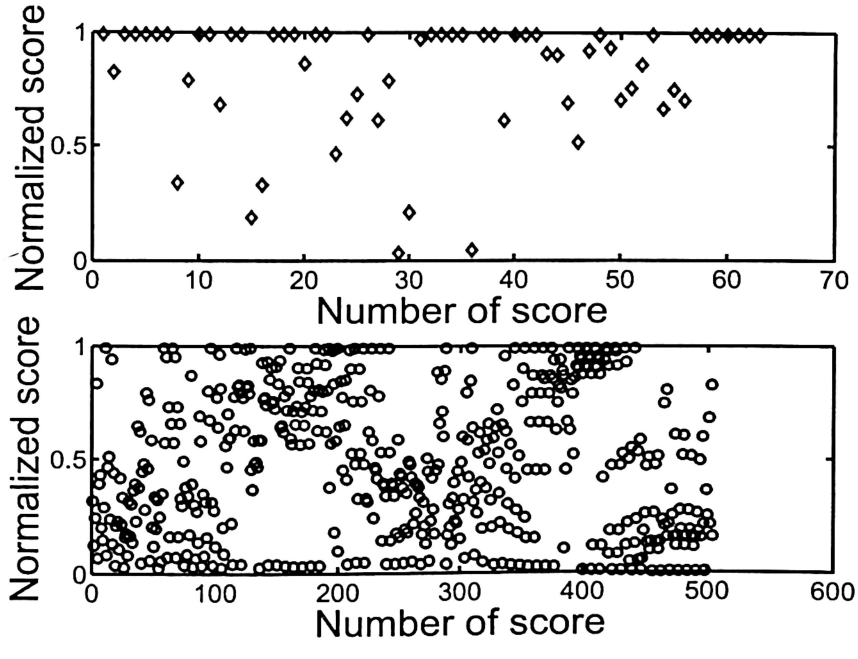


Figure 3.9: Genuine (diamond symbol) and Imposter (circular symbol) score of Cepstral coefficients

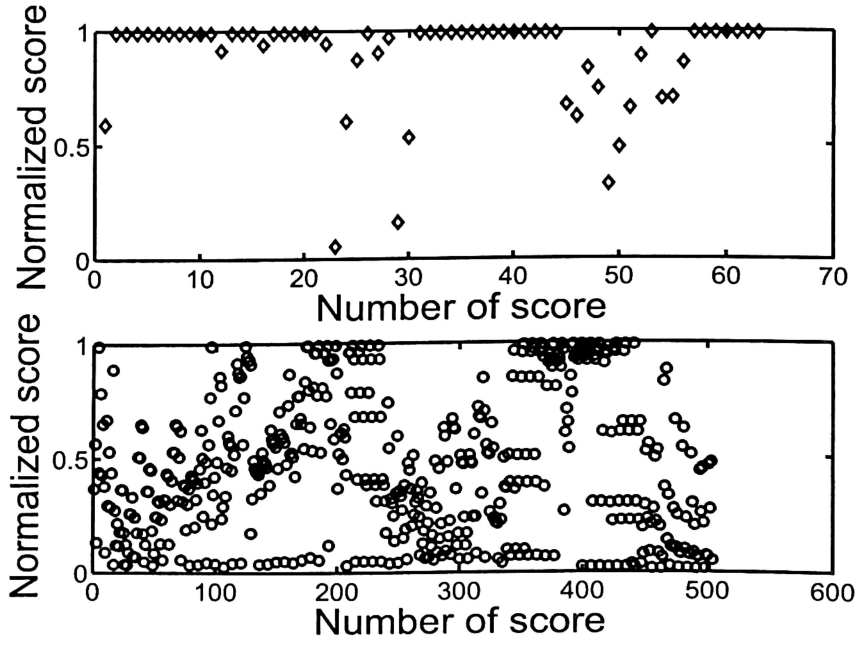


Figure 3.10: Genuine (diamond symbol) and Imposter (circular symbol) score of Arc-sin reflection coefficients

tion based trained models show maximum average recognition for all emotion testing utterances for LPCC, LAR, MFCC and RC. This result shows that Question emotion based speaker models may contain more speaker information than others or it may contain speaker information which are invariant in other emotions. These results suggest that Question speech data may require more attention from the

TH-1853_02610201

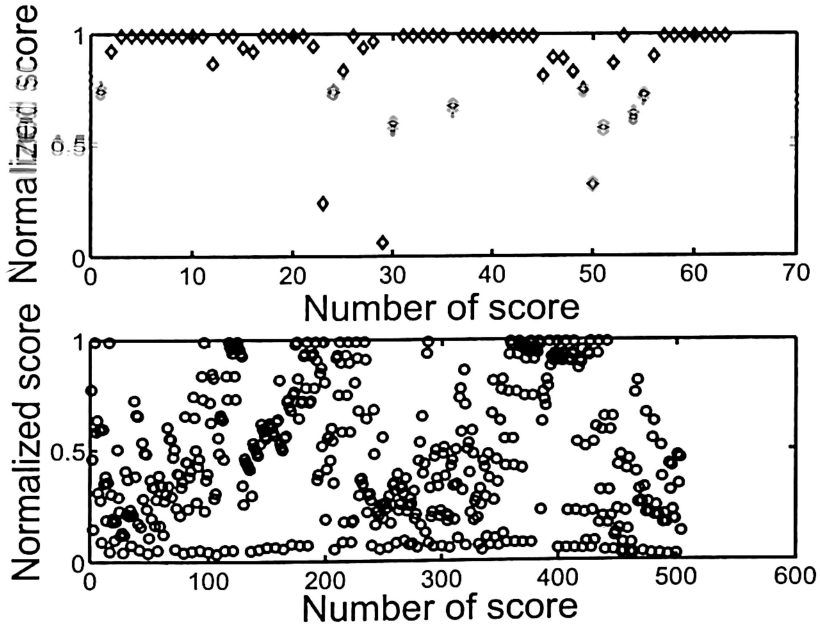


Figure 3.11: Genuine (diamond symbol) and Imposter (circular symbol) score of Reflection coefficients

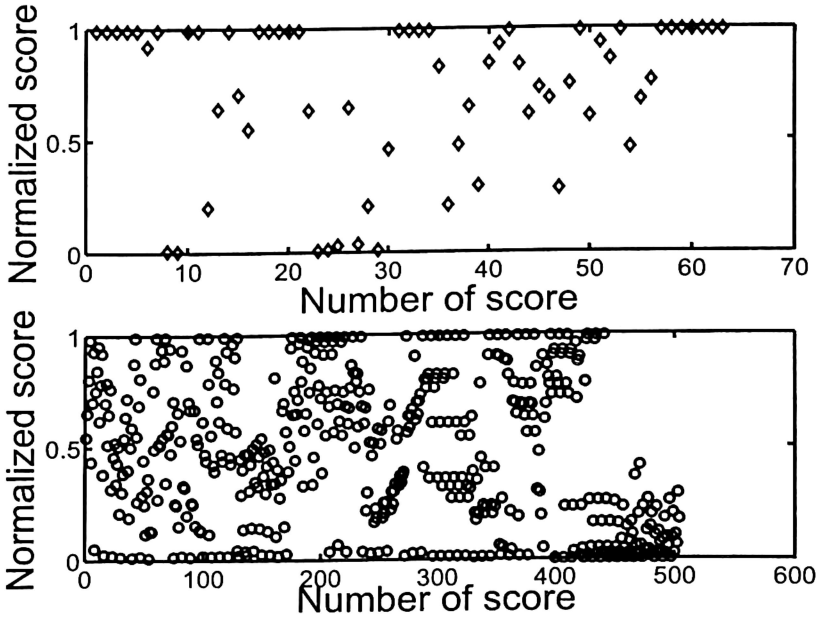


Figure 3.12: Genuine (diamond symbol) and Imposter (circular symbol) score of Linear prediction coefficients

point of view of speaker information.

3.2.2 Gaussian Mixture Model (GMM) based classifier

The discussions in this subsection will focus on GMM speaker model and speaker recognition under stressed conditions. First, it is observed that probability density function characteristics of the features

TH-1853_02610201

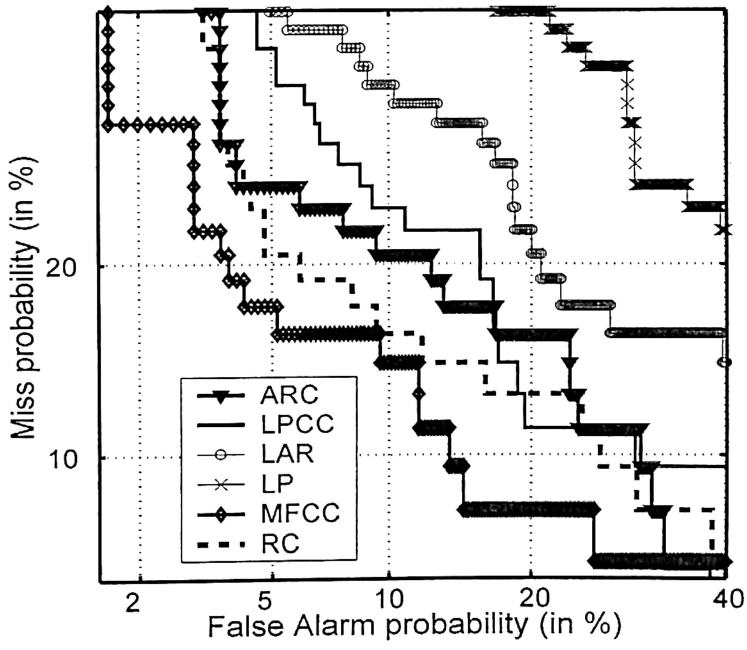


Figure 3.13: DET curve for different features under neutral condition

Table 3.7: Speaker recognition results under non pre-emphasized (Neutral emotion) condition

SI Identification rate in % (pre-emphasized)	SI Identification rate in % (non pre-emphasized)	SV EER in % (pre-emphasized)	SV EER in % (non pre-emphasized)
RC(82.53)	LAR(80.95)	MFCC (10.52)	RC(13.59)
MFCC(80.95)	RC(80.95)	RC(12.50)	ARC(14.58)
LAR(79.36)	ARC(80.95)	ARC(14.68)	MFCC(14.68)
ARC(79.36)	MFCC(79.36)	LPCC(15.28)	LAR(13.69)
LPCC(71.42)	LPCC(76.19)	LAR(19.94)	LPCC(21.13)
LP(58.73)	LP(65.07)	LP(27.68)	LP (36.41)

follow closely Gaussian distribution. Second, the model naturally arises from a probabilistic modelling of the underlying sounds. GMM [16] basically models the feature vectors of the speaker as Gaussian densities. Table 3.12 shows the GMM based speaker identification results for pre-emphasized speech.

TH-1853_02610201

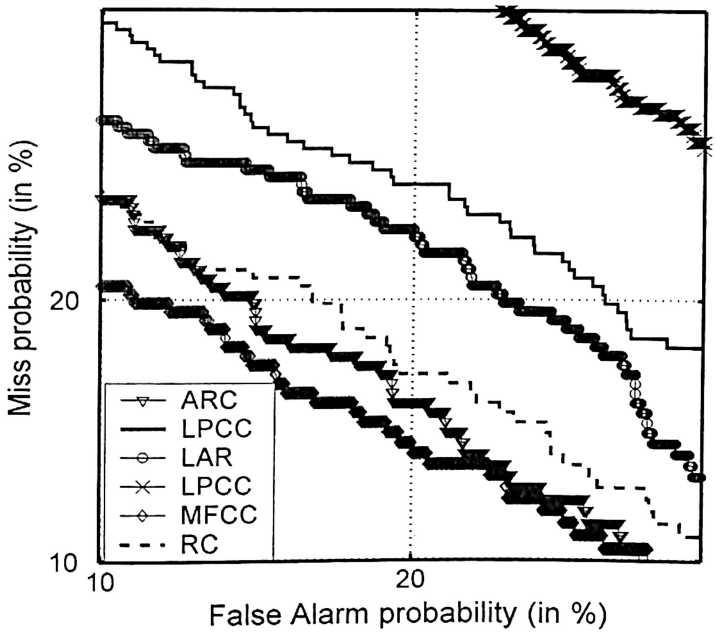


Figure 3.14: DET curve for different features under stressed condition

Table 3.8: Speaker recognition results under stressed condition with non pre-emphasized (Neutral emotion) speaker models

SI Identification rate in % (pre-emphasized)	SI Identification rate in % (non pre-emphasized)	SV EER in % (pre-emphasized)	SV EER in % (non pre-emphasized)
ARC (54.75)	RC (53.57)	MFCC(14.06)	MFCC(12.95)
MFCC (53.96)	ARC (51.58)	RC(16.84)	ARC(16.64)
RC (53.96)	LAR (50.79)	ARC(16.47)	RC(18.55)
LAR (50.79)	MFCC (47.22)	LAR(19.20)	LAR(19.62)
LPCC (48.01)	LPCC (46.43)	LPCC(21.58)	LPCC(22.35)
LP (43.64)	LP (40.87)	LP(29.54)	LP(39.73)

A 16 order GMM is considered to model each feature. Speaker models are prepared by the features extracted from Neutral speech data. Each column of this table shows speaker identification rate for each reference emotion testing utterance. Pre-emphasized speech gives more recognition for Neutral testing

TH-1853_02610201

Table 3.9: Speaker identification rate (%) for different emotions using MFCC

Training (256 size codebook)	**Testing**				
	Angry	Lombard	Neutral	Question	Average
12c(Pre-emphasized)	30.15	49.20	76.19	60.31	53.96
20c(Pre-emphasized)	33.01	51.57	80.51	62.90	56.99
12c(Non Pre-emphasized)	22.22	44.44	80.95	50.38	49.99
20c(Non Pre-emphasized)	23.25	47.95	81.88	52.49	51.39

utterances as the trained model is developed with the Neutral speech data. The recognition results are higher for the Neutral test utterances in case of GMM compared to those in case of VQ based classifier shown in third column of Table 3.5. The average recognition results are lower in case of VQ compared to those in case of GMM based classifier shown in the last column of Table 3.5. This may be due to the fact that GMM is based on mean and variance of the features where as VQ is based on the mean value.

3.3 Relative Ranking of the Features

In this chapter, six different speech features, LP coefficients, LPCC, ARC, RC, LAR and MFCC, are used for evaluation of speaker recognition under stressed condition. Different statistical feature evaluation techniques, such as probability density characteristics, F-ratio test, multi-dimensional F-ratio test (Divergence), Kolmogorov-Smirnov test are used to assess the performances of the speech features. Vector quantization and Gaussian mixture model are used to assess the recognition performance. Speech data from SUSAS database corresponding to four different emotions of stressed conditions, Angry, Lombard, Question and Neutral, are used for analysis of speaker recognition under stressed condition. One of the goals of the analysis is to select features which are robust to stressed conditions. Relative ranking of the performance of different speech features will serve this purpose. Statistical analysis tools such as F-ratio test and K-S test provide insight into relative effectiveness of the coefficients of a feature in discriminating the speakers. F-ratio test helps identify the feature coefficients which are relatively less effective in classification. It is difficult to evaluate the relative performance of the features based on their statistical analysis. This is evident from the differences in observations between the relative average F-ratio values and the relative speaker recognition values of the features. It is observed from the first column of Table 3.1 that LAR feature has maximum average F-ratio value and MFCC feature has minimum average F-ratio value for pre-emphasized speech. But, speaker recogni-

TH-1853_02610201

tion results tabulated in Table 3.8 shows that ARC and MFCC features have best speaker identification and speaker verification results respectively. These observations indicate that it may not be correct to evaluate the relative performance of the features based upon the average F-ratio values or based upon the other statistical tools used for analysis of the features. As speaker recognition is of primary interest in this work, the ranking of the features can be done according to their recognition performances. Hence, for relative ranking of the features, the recognition performances are considered.

In this chapter, four types of speaker recognition experiments, VQ based speaker identification (SI), VQ based speaker verification (SV), GMM based SI and GMM based SV, are performed. The speaker recognition analysis are done based on pre-emphasized speech and non pre-emphasized speech. Mostly the above analysis shows that pre-emphasized speech gives more recognition than non pre-emphasized speech for speaker recognition under stressed condition. So, we can rank the speech features for speaker recognition under stressed condition using these pre-emphasized speech. Table 3.13 shows the performance of different features when evaluated by the four different stress recognition experiments. The recognition results of the six speech features are shown in four columns in descending order of their performances. Though the relative performance of the features is different in the four columns, MFCC, RC and ARC features have better performance compared to the other features. Similarly, LAR, LPCC and LP features show poor recognition results compared to other features in all the four recognition experiments. It is observed that the results of four speaker recognition experiments are different from each other. Under these circumstances, it is difficult to rank the features for their effectiveness in speaker recognition under stressed condition. We suggest the following approach for ranking the features.

For ranking the speech features, the probabilities of their positions in the four columns in Table 3.13 are evaluated. Each feature appears once in each column. Each position in a row is assigned a probability value of 0.25 for a feature if it appears. ARC and RC features appear once in the first row. So, they are assigned a probability value of 0.25 for first ranking. MFCC appears twice in the first row. So, it is assigned a probability value of 0.5 for the second ranking. Similarly, probability values of the features according to their positions are estimated and are shown in Table 3.14. Each row in this table shows the ranking probabilities of the features. Final ranking of features is decided based on the cumulative probability. Cumulative probability values are shown in Table 3.15. Cumulative probabilities of MFCC, RC and ARC features attain the value of one by the third row. Hence, these three features qualify for the top three rankings. In the second row, cumulative probability values of MFCC and RC features are same which is equal to 0.75 and the ARC feature has a value of 0.5. So, ARC feature is ranked third. MFCC and RC features have cumulative probability values of 0.5 and

TH-1853_02610201

0.25 respectively in the first row. So, MFCC feature is ranked first and RC feature is ranked second. LAR and LPCC features attain cumulative probability values of one by fifth row. Based on their values in the fourth row, LAR and LPCC features are ranked as the fourth and fifth features respectively. LP feature is ranked as the sixth feature. MFCC feature is ranked as the best feature followed by RC, ARC, LAR, LPCC and LP features in descending order of their performances in speaker recognition under stressed condition.

TH-1853_02610201

Table 3.10: Speaker identification rate (%) for non pre-emphasized speech

Feature: ARC					
Training (256 size codebook)	**Testing**				
	Angry	Lombard	Neutral	Question	Average
Angry	71.42	44.44	46.03	33.33	48.805
Lombard	33.33	92.06	44.44	31.74	50.3925
Neutral	38.09	46.03	82.53	65.07	57.93
Question	31.74	46.03	79.36	80.95	59.52

Feature: LPCC					
Training (256 size codebook)	**Testing**				
	Angry	Lombard	Neutral	Question	Average
Angry	61.9	42.85	36.5	28.57	42.455
Lombard	23.8	82.53	55.55	36.5	49.595
Neutral	34.92	46.03	87.3	61.9	57.5375
Question	26.98	42.85	66.66	65.07	50.39

Feature: LAR					
Training (256 size codebook)	**Testing**				
	Angry	Lombard	Neutral	Question	Average
Angry	74.6	42.85	44.44	34.92	49.2025
Lombard	31.74	90.47	46.03	28.57	49.2025
Neutral	25.39	39.68	84.12	60.31	52.375
Question	25.39	41.26	73.01	77.77	54.3575

Feature: LP					
Training (256 size codebook)	**Testing**				
	Angry	Lombard	Neutral	Question	Average
Angry	55.55	34.92	33.33	31.74	38.885
Lombard	25.39	74.6	49.2	28.57	44.44
Neutral	31.74	53.96	65.07	44.44	48.8025
Question	25.39	31.74	46.03	63.49	41.6625

Feature: MFCC					
Training (256 size codebook)	**Testing**				
	Angry	Lombard	Neutral	Question	Average
Angry	63.49	38.09	42.85	28.57	43.25
Lombard	17.46	85.71	46.03	33.33	45.6325
Neutral	22.22	44.44	80.95	52.38	49.9975
Question	28.57	39.68	69.84	79.36	54.3625

Feature: RC					
Training (256 size codebook)	**Testing**				
	Angry	Lombard	Neutral	Question	Average
Angry	71.42	49.2	46.03	31.74	49.5975
Lombard	36.5	88.88	42.85	31.74	49.9925
Neutral	39.68	44.44	84.12	60.31	57.1375
Question	30.15	41.26	73.01	71.42	53.96

TH-1853_02610201

Table 3.11: Speaker identification rate (%) for pre-emphasized speech

Feature: ARC					
Training (256 size codebook)	**Testing**				
	Angry	Lombard	Neutral	Question	Average
Angry	71.42	52.38	53.96	38.09	53.9625
Lombard	31.74	88.88	57.14	39.68	54.36
Neutral	38.09	49.2	79.36	66.66	58.3275
Question	33.33	47.61	68.25	76.19	56.345
Feature: LPCC					
Training (256 size codebook)	**Testing**				
	Angry	Lombard	Neutral	Question	Average
Angry	55.55	44.44	36.5	23.8	40.0725
Lombard	25.39	82.53	55.55	39.68	50.7875
Neutral	26.98	46.03	77.77	55.55	51.5825
Question	31.74	47.61	69.84	68.25	54.36
Feature: LAR					
Training (256 size codebook)	**Testing**				
	Angry	Lombard	Neutral	Question	Average
Angry	66.66	49.2	52.38	36.5	51.185
Lombard	34.92	85.71	49.2	39.68	52.3775
Neutral	31.74	42.85	82.53	63.49	55.1525
Question	33.33	44.44	79.36	73.01	57.535
Feature: LP					
Training (256 size codebook)	**Testing**				
	Angry	Lombard	Neutral	Question	Average
Angry	50.79	46.03	46.03	31.74	43.6475
Lombard	36.5	87.3	49.2	28.57	50.3925
Neutral	28.57	50.79	68.25	47.61	48.805
Question	31.74	36.5	58.73	65.07	48.01
Feature: MFCC					
Training (256 size codebook)	**Testing**				
	Angry	Lombard	Neutral	Question	Average
Angry	71.42	41.26	41.26	36.5	47.61
Lombard	20.63	82.53	44.44	39.68	46.82
Neutral	30.15	49.2	76.19	60.31	53.9625
Question	25.39	49.2	73.01	73.01	55.1525
Feature: RC					
Training (256 size codebook)	**Testing**				
	Angry	Lombard	Neutral	Question	Average
Angry	71.42	49.2	50.79	46.03	54.36
Lombard	28.57	88.88	60.31	33.33	52.7725
Neutral	36.5	52.38	84.12	60.31	58.3275
Question	34.92	44.44	73.01	82.53	58.725

TH-1853_02610201

Table 3.12: Speaker identification rate using GMM classifier for pre-emphasized speech (Training: Neutral)

Features	Speaker identification rate (%)				
	Emotion of testing utterances				Average
	Angry	Lombard	Neutral	Question	
RC	30.15	49.2	88.88	52.38	55.15
ARC	30.15	50.79	85.71	52.38	54.75
MFCC	26.98	46.03	90.47	50.79	53.56
LPCC	20.63	52.38	84.12	55.55	53.17
LAR	30.15	39.68	82.53	52.38	51.18
LP	33.33	42.85	71.42	50.79	49.59

Table 3.13: Feature analysis for speaker recognition under stressed condition (pre-emphasized)

Rank	VQ based Speaker identification rate(%)	VQ based Speaker verification (EER in %)	GMM based speaker identification rate(%)	GMM based speaker verification (EER in %)
1	ARC (54.75)	MFCC(14.06)	RC(55.15)	MFCC (9.7)
2	MFCC (53.96)	RC(16.84)	ARC(54.75)	RC(13.39)
3	RC (53.96)	ARC(16.47)	MFCC(53.56)	ARC(13.79)
4	LAR (50.79)	LAR(19.20)	LPCC(53.17)	LAR (14.83)
5	LPCC (48.01)	LPCC(21.58)	LAR(51.18)	LPCC(16.82)
6	LP (43.64)	LP(29.54)	LP(49.59)	LP(21.78)

TH-1853_02610201

Table 3.14: Position probability of six speech features for speaker recognition under stressed condition

Position	MFCC	LPCC	LAR	ARC	RC	LP
1	0.5	0	0	0.25	0.25	0
2	0.25	0	0	0.25	0.5	0
3	0.25	0	0	0.5	0.25	0
4	0	0.25	0.75	0	0	0
5	0	0.75	0.25	0	0	0
6	0	0	0	0	0	1

Table 3.15: Position probability of six speech features for speaker recognition under stressed condition

Position	MFCC	RC	ARC	LAR	LPCC	LP
1	0.5	0.25	0.25	0	0	0
2	0.75	0.75	0.5	0	0	0
3	1	1	1	0	0	0
4				0.75	0.25	0
5				1	1	0
6						1

TH-1853_02610201

Chapter 4

Speaker Information Analysis using Sinusoidal Model Features

Sinusoidal model based speech coder is well established for 4 kHz bandwidth transmission lines [118]. There is little reported work on sinusoidal model based speaker information analysis. Sinusoidal model can produce three sets of features. They are Amplitude, Frequency and Phase features. Assuming that the information in these three features may be different, these features may be combined for improvement in speaker recognition. Also, sinusoidal features preserve the formant information as peaks in the spectrum are used for estimation of amplitude, frequency and phase features. From the signal reconstruction point of view, sinusoidal model features have advantage over the mel-frequency cepstral coefficients (MFCC) which are widely accepted features for speaker recognition. Sinusoidal modeling has been applied to different problems in the area of speech science and has advantages over linear predictive modeling and short time Fourier transform [119], [120], [121]. Co-channel talker interference suppression is achieved by synthesizing a waveform based on a sinusoidal representation of speech [122]. Sinusoids with time-varying frequencies (and amplitudes) enable a more accurate representation of non stationary speech frames [123]. High reconstructed speech quality can be achieved by accurate estimation of component frequencies (peaks) and elimination of side lobe interference effects [2].

This chapter presents speaker recognition using sinusoidal model features. Frequency, Amplitude and Phase features of the sinusoidal model are analyzed and used as input features to a speaker recognition system. Speaker recognition performance of the three sinusoidal model features is evaluated under stressed speech conditions. In Section 4.1, sinusoidal modeling of speech is presented. Statistical analysis of sinusoidal model features is discussed in Section 4.2. Speaker recognition performance

TH-1853_02610201

sinusoidal model features under stressed speech condition is discussed in Section 4.3. A novel approach of using both MFCC and sinusoidal model features to improve the speaker recognition performance is proposed and the results are presented in Section 4.4.

4.1 Sinusoidal Modeling of Speech Signal

Sinusoidal model parameter extraction algorithm shows that the Amplitudes, Frequencies and Phases of the sine waves can be obtained from the high-resolution short-time Fourier transform (STFT) by locating the peaks of the associated magnitude function [125]. Friedlander et. al. [126] presents a maximum likelihood algorithm for estimating the Phase and Amplitude features. Speech signal can be viewed as the result of passing a glottal excitation waveform through a time varying linear filter which models the resonant characteristics of the vocal tract. In sinusoidal modeling [125], the non-stationary speech signal is divided into number of stationary frames. The speech frame $\tilde{s}[n]$ in terms of overlap-add sinusoidal model can be represented as,

$$\tilde{s}[n] = \sigma^{(k)}[n] \sum_{k=1}^{M} \tilde{w}(n - kN)\tilde{s}^{(k)}(n - kN) \tag{4.1.1}$$

$$\sum_{k=1}^{M} \tilde{w}(n - kN) = 1 \tag{4.1.2}$$

where M is number of frames, k is frame index and N is length of frame. Each frame is visualized as made up of sum of L sinusoids and features (Frequency, Amplitude and Phase) of each sinusoid are estimated as follows.

$$\tilde{s}^{(k)}[n] = \sum_{j=1}^{L} A_j^{(k)} \cos(2\pi f_j^{(k)} \frac{n}{F_s} + \phi_j^{(k)}) \tag{4.1.3}$$

where F_s is the sampling frequency of the speech signal, $s[n]$. $A_j^{(k)}$, $\phi_j^{(k)}$ and $f_j^{(k)}$ represents Amplitude, Phase and Frequency of the j^{th} sinusoid of k^{th} frame respectively. Block diagram of sinusoidal model feature extraction method is shown in figure 4.1. The speech signal is divided into number of frames (frame with the length of 20 ms and an overlap of 10 ms). To remove the silent periods of speech signal, frames at beginning and end are removed if they contain energy less than 10% of the total energy. Then Hamming window is applied to remove the discontinuities at the end of each frame. The amplitude spectrum is obtained from FFT of windowed signal, from which most significant peaks are selected as coefficients or components of Amplitude feature. The amplitude peaks are mapped over the frequency track and corresponding frequency values are selected as coefficients of Frequency

TH-1853_02610201

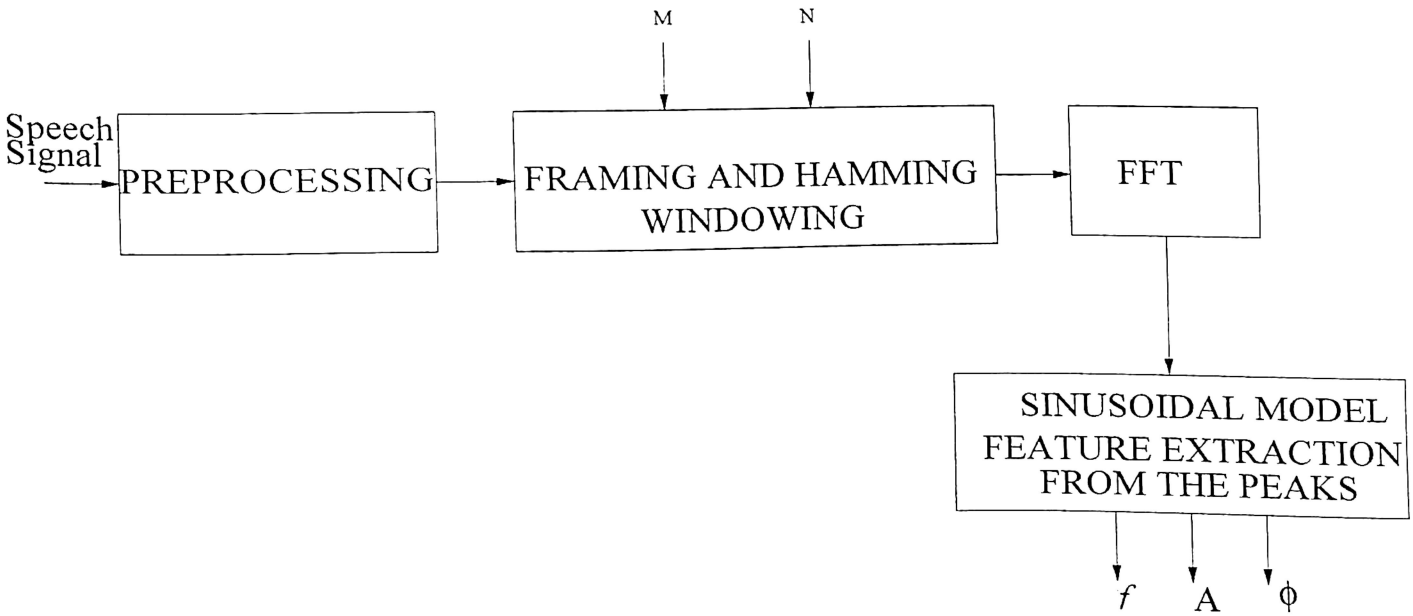


Figure 4.1: Sinusoidal model feature extraction

feature. Similarly coefficients of Phase feature are estimated corresponding to those peaks. All the feature vectors are arranged according to the frequency track. These Amplitude, Frequency and Phase features constitute three sets of features of sinusoidal model. They are named as sinusoidal amplitude feature (SAF), sinusoidal frequency feature (SFF) and sinusoidal phase feature (SPF). For evaluation of speaker recognition performance, logarithm of the peak values are used as coefficient values of SAF.

Sinusoidal model is tested by reconstruction of voiced and unvoiced speech signals. For this evaluation, the extracted Amplitude, Frequency and Phase features are used for reconstruction. Figure 4.2 shows a voiced speech signal and the reconstructed signal for a model order of 12. Figure 4.3 shows a unvoiced speech signal and the reconstructed signal. It is observed that the voiced speech signal is reproduced better compared to the unvoiced speech signal. The unvoiced speech signal may require a higher order model. The mean square error between the original speech signal and the reconstructed speech signal is calculated to for selection of model order. Figure 4.4 shows the variation of the mean square error with the model order for a voiced speech. It is observed that the error energy decreases with the increase of model order. The error energy is less beyond the 12th order. A model order of ten or higher for sinusoidal model can be used for evaluation of speaker recognition [81, 82, 127]. For speaker discrimination analysis in the next subsection, a 12^{th}-order model is used.

TH-1853_02610201

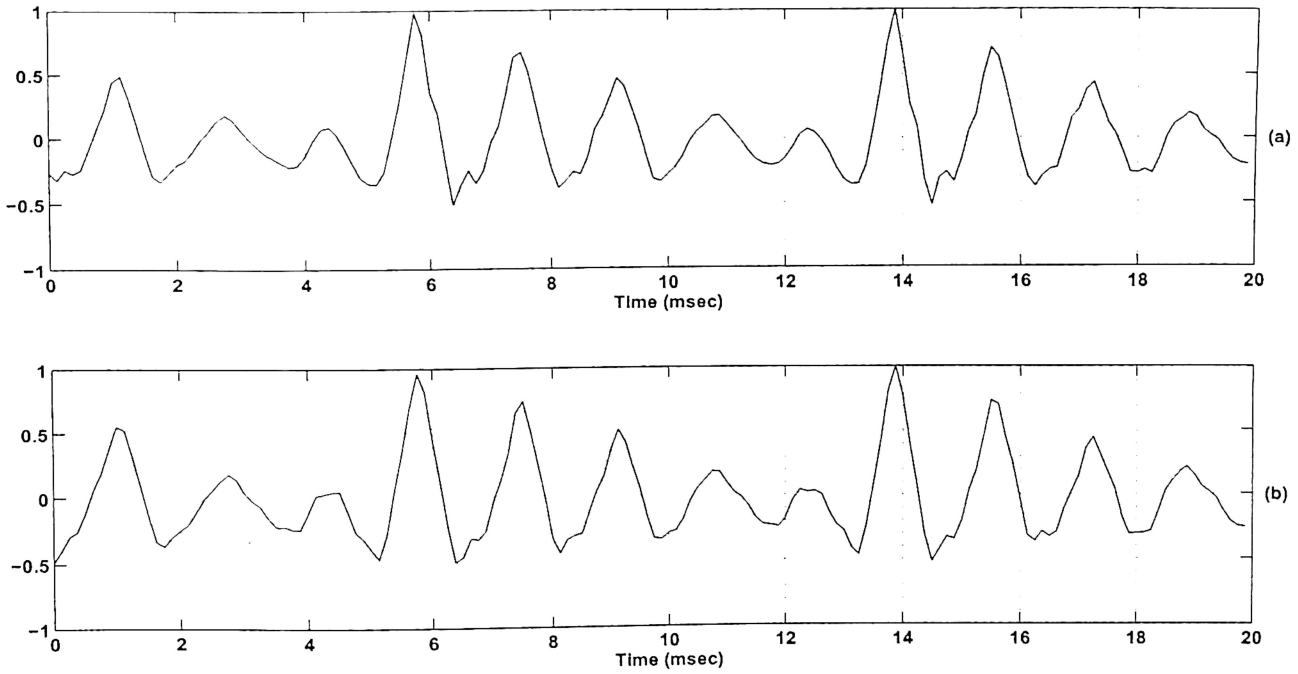


Figure 4.2: (a) Voiced speech signal, (b) Reconstructed voiced speech signal

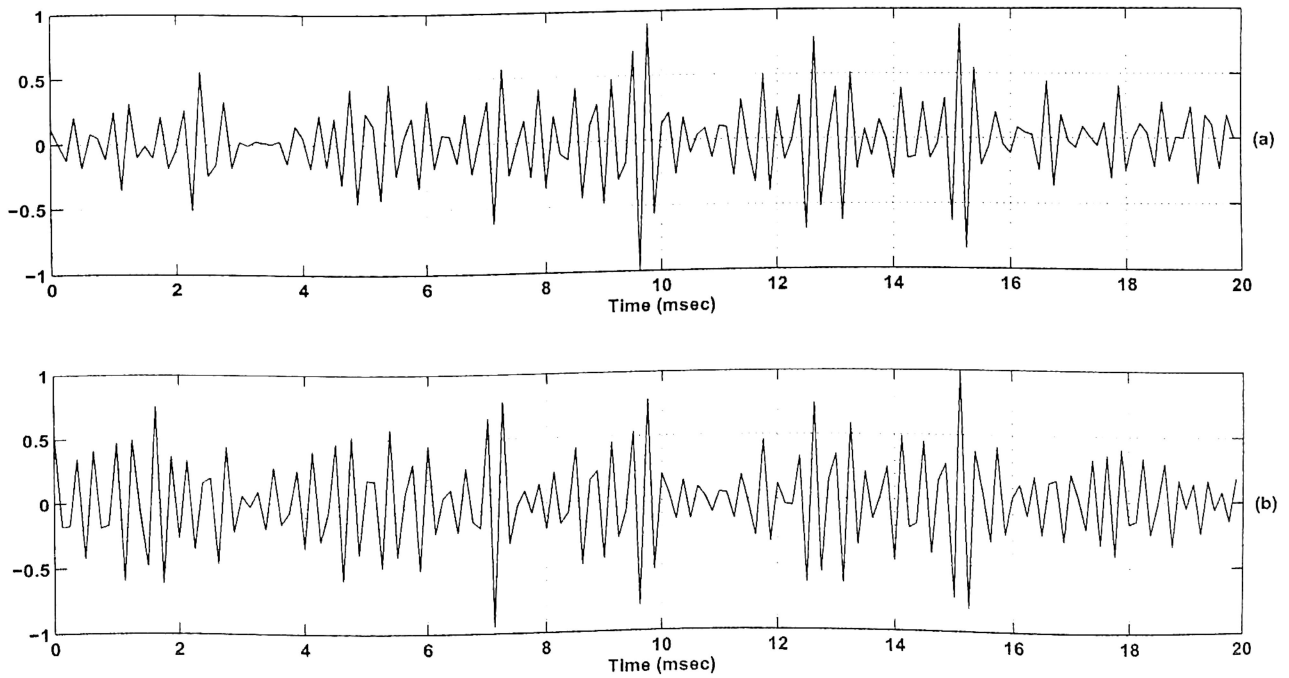


Figure 4.3: (a) Unvoiced speech signal, (b) Reconstructed unvoiced speech signal

TH-1853_02610201

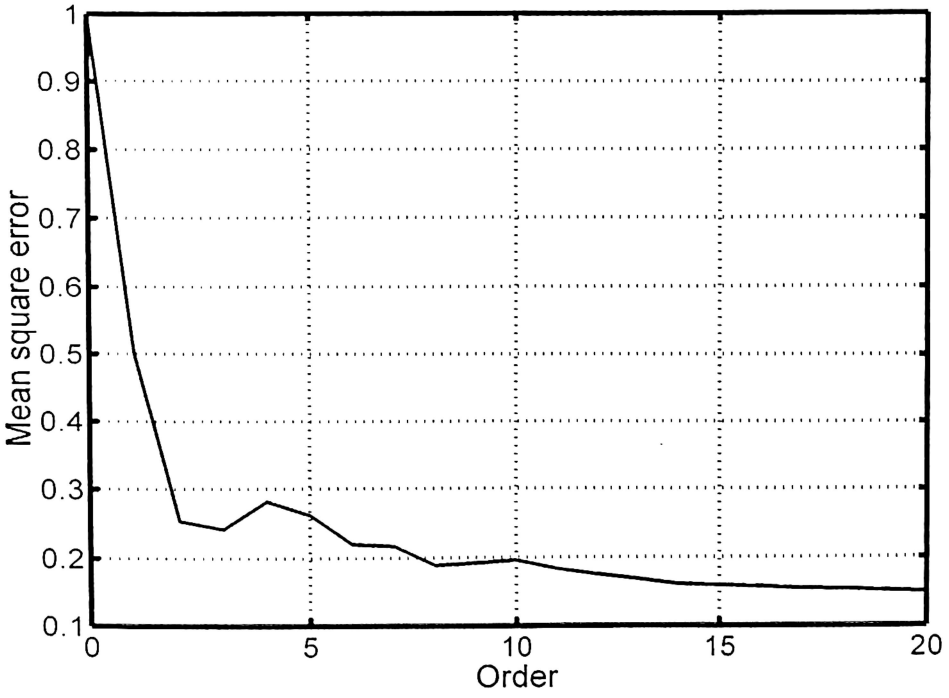


Figure 4.4: Mean square error versus order for a voiced speech signal

4.1.1 Speaker discrimination with sinusoidal features

Sinusoidal model features are expected to carry speaker specific information. Amplitude contour plot is a method to show the difference between the speakers. Fig. 4.5 shows A_5 Amplitude contours for five different speakers. Forty utterances and sixty frames from each utterance are considered for this contour plot. The average values of A_5 coefficient over the forty utterances for each speaker is plotted as A_5 Amplitude contour. It is observed that the A_5 Amplitude contour is different for different speakers. The average values of A_5 Amplitudes of five speakers are 0.0079, 0.0077, 0.0082, 0.0160 and 0.0096. Similar results are also observed for other amplitude contours A_1 to A_{12}. This shows that Amplitude feature can be used for speaker recognition.

Similar to Amplitude contour plot, we have analyzed the frequency contour. Fig. 4.6 shows the f_5 contours of the five different speakers. These contours which can be viewed as a time varying representations of f_5 frequency, show significant variation between different speakers. The lowest frequency gives the near estimate of fundamental frequency of that particular speaker. The average values of f_5 frequencies of the same speech uttered by five different speakers are 0.9720kHz, 0.9419kHz, 0.9082kHz, 0.8304kHz and 0.9113KHz. We observe similar characteristic differences between the frequency contours of different speakers for all the ten frequencies. For investigating the statistical

TH-1853_02610201

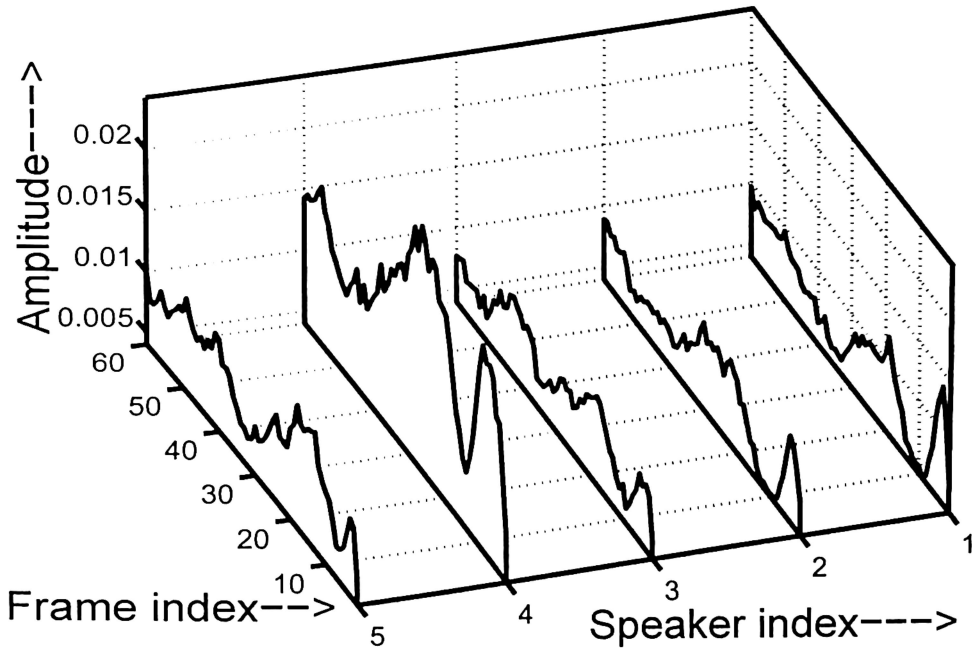


Figure 4.5: Amplitude contour

characteristics of these frequency contours, their probability densities are evaluated. The probability densities for different frequencies are estimated by using training utterances from each of the speakers. It shows that the frequency contour can characterize the speakers.

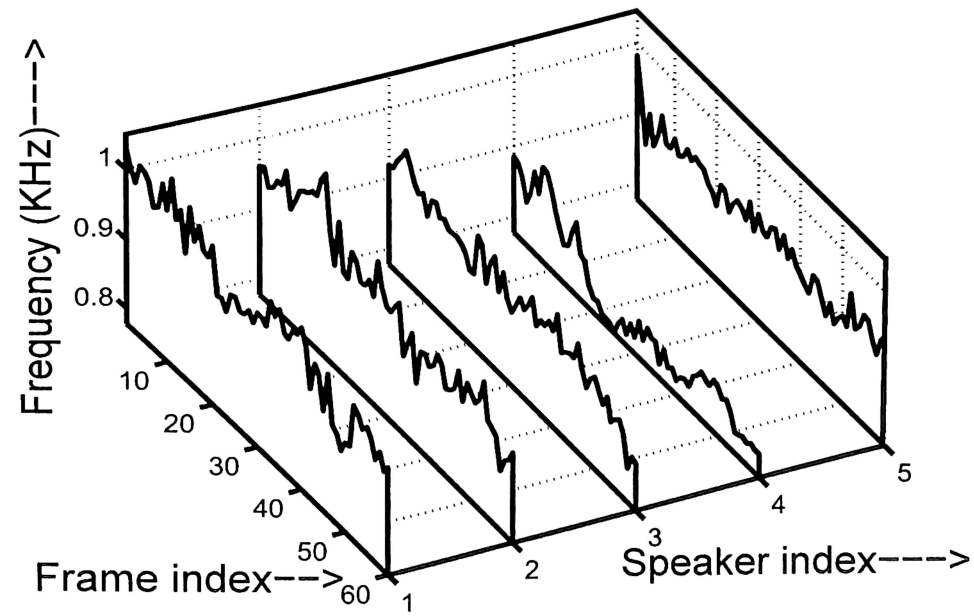


Figure 4.6: Frequency contour

The probability density function (pdf) characteristic of a coefficient of four speakers can be es-

TH-1853_02610201

timated for comparison. The probability density characteristics will indicate that the coefficient is suitable for speaker recognition if the peaks of the probability densities are placed at different positions for different speakers. The probability density function characteristics of sinusoidal amplitude feature (SAF), sinusoidal frequency feature (SFF) and sinusoidal phase feature (SPF) for four speakers are shown in Figures 4.7, 4.8, 4.9. Fifth coefficient of the three features are chosen for the above probability density function (pdf) characteristics. It is observed that the pdf characteristics closely resemble Gaussian density characteristics. The A_5 coefficient of SAF pdf characteristics of four speakers have peaks placed at different positions. Also, the variances are different for four speakers. This suggests that the this SAF coefficient is suitable to capture the speaker discriminating information. Similar results are observed with the other coefficients of the Amplitude feature. From Fig. 4.8, it is observed that the f_5 coefficient pdf characteristics of four speakers have peaks placed at different positions and they have different variance values. Similar properties are observed with the other coefficients of the SFF. The pdf characteristics of ϕ_5 coefficient of the SPF shown in Fig. 4.9 do not show much difference in their peak positions and variance values for the four speakers. This suggests that this coefficient of the SPF may not show much speaker discriminating capability. Similar results are obtained for other coefficients of the Phase feature. These results suggest that SAF and SFF feature have better speaker discriminating capabilities compared to SPF features. Speaker recognition results in the following sections prove this point.

4.2 Statistical Analysis of Sinusoidal Model Features

In this section, sinusoidal model features are analyzed using the statistical tests. In the previous chapter, statistical analysis of six speech features is carried out to find out their speaker discriminating capabilities. For all the six features, twelfth order model has been considered in the previous chapter. To analyze the sinusoidal features and compare them with the earlier six features, a twelfth order sinusoidal model is considered in this analysis. Fig. 4.10 shows F-ratio values for the three sinusoidal model features. F-ratios evaluated between nine speakers are plotted with solid lines and F-ratios evaluated between four emotion data are plotted with dashed lines. It is observed that 12^{th} coefficient of SAF feature has maximum values for both the speaker and emotion cases. This suggests that this coefficient is the most important coefficient from both speaker and stressed recognition point of view. Similarly twelfth coefficients is the most significant SFF coefficient from speaker and emotion recognition. F-ratio values for the sinusoidal phase feature (SPF) are less compared the F-ratio values for the SAF and SFF. For SPF,

TH-1853_02610201

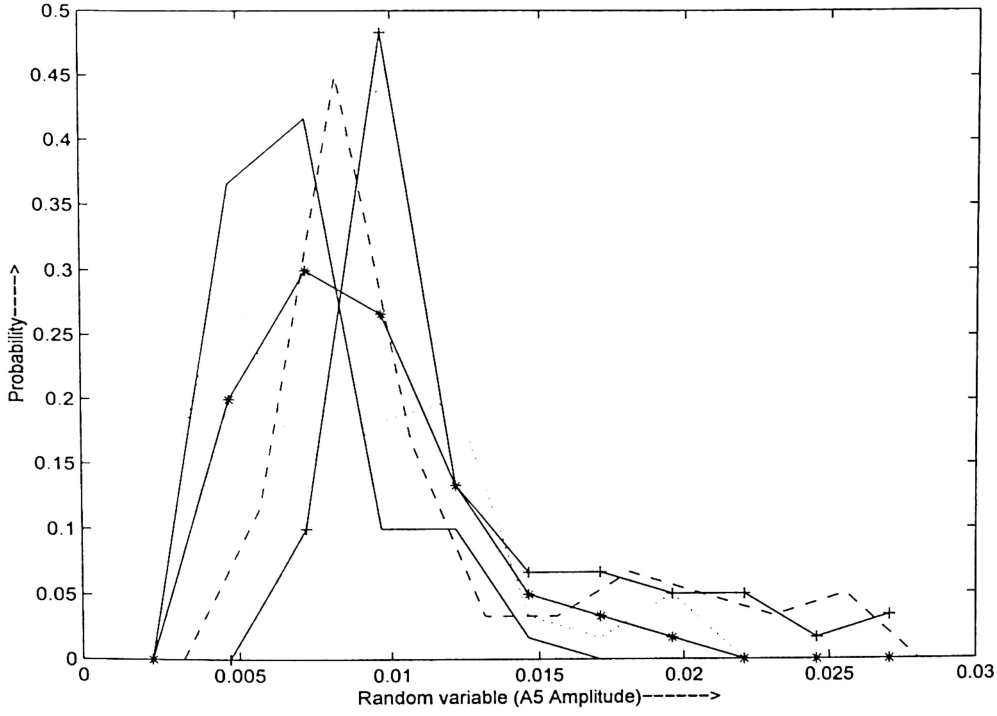


Figure 4.7: Probability density function of amplitude feature for four speakers

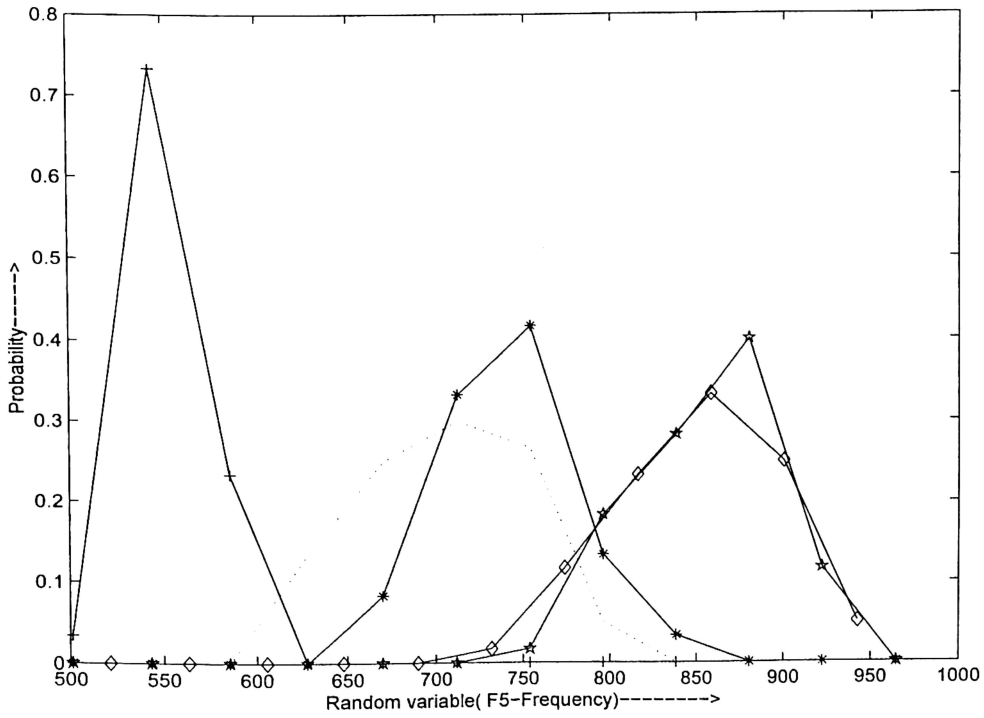


Figure 4.8: Probability density function of frequency feature for four speakers

TH-1853_02610201

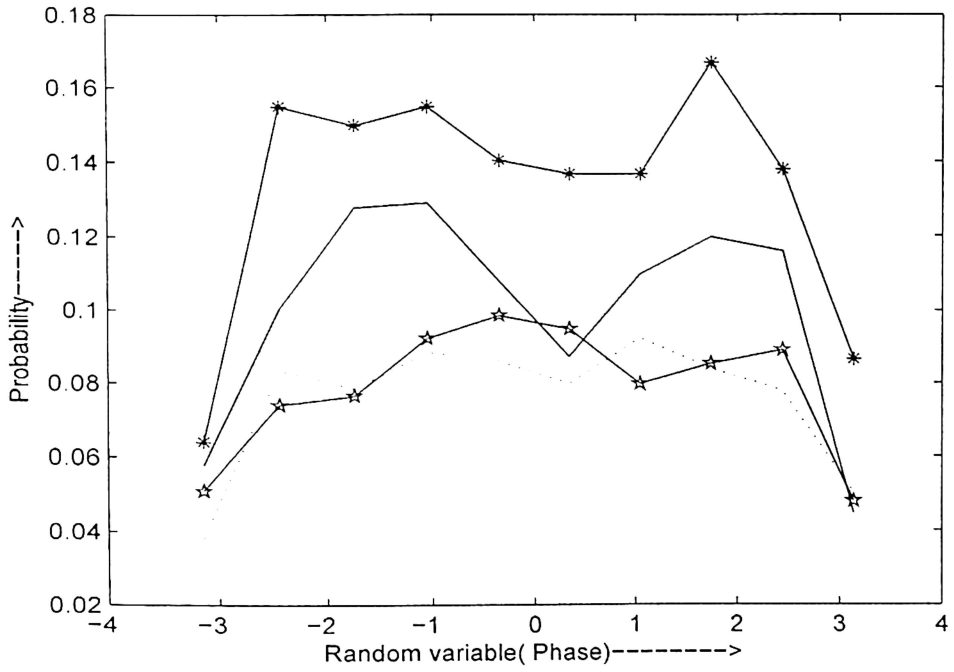


Figure 4.9: Probability density function of phase features for four speakers

first coefficient is most significant for speaker recognition whereas sixth coefficient carries maximum emotion discriminating information.

Average F-ratio, average KS-test distance and Divergence values of sinusoidal model features are tabulated in Table 4.1. KS-test and F-ratio test are calculated for each coefficient separately and then average of these values are estimated. Multi-dimensional F-ratio test or divergence is calculated considering the feature vectors instead of individual coefficients. For the stressed speech analysis, 42 utterances from SUSAS database are considered for each emotion. Four emotions such as Angry, Lombard, Neutral and Question are used for this analysis. Nine speakers are considered for evaluation of these statistical values. It is observed that SAF has the highest average value for all the three statistical parameters followed by the values for SFF and SPF. Hence, sinusoidal amplitude feature (SAF) may perform best followed by SFF and SPF for speaker recognition.

4.3 Speaker Recognition with Sinusoidal Model

For extraction of sinusoidal features, the spectrum is estimated for each frame of the speech signal (after windowing) by using 256-point FFT. From the magnitude spectrum, significant peaks are estimated

TH-1853_02610201

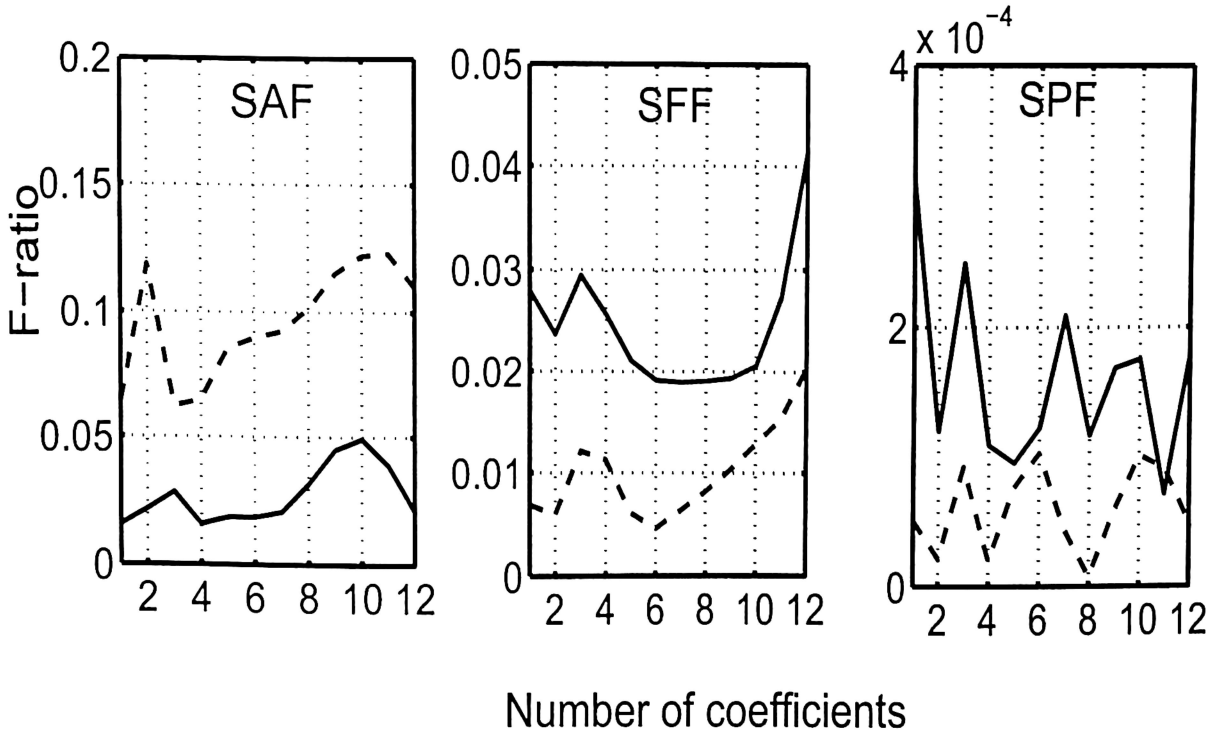


Figure 4.10: F-ratios of sinusoidal features. Solid line shows F-ratios when evaluated between the speakers and dashed line shows F-ratios when evaluated between the four emotions.

Table 4.1: Average F-ratio, average KS-test distance and Divergence values of sinusoidal model features under stressed condition

Rank	Average F-ratio	Divergence	Average KS-test distance
1	SAF(0.0230)	SAF(3.101)	SAF(0.094)
2	SFF(0.0089)	SFF(1.749)	SFF(0.091)
3	SPF(0.0001)	SPF(0.0024)	SPF(0.001)

at which the slope of the magnitude spectrum changes from positive to negative. Frequencies, f_j, are estimated by arranging the frequency values corresponding to the peaks in an ascending order. The amplitudes corresponding to the above frequencies are named as A_j amplitudes. From the phase spectrum, the phases, ϕ_j, corresponding to the above frequencies are chosen as coefficients of sinusoidal phase feature (SPF). This method is considered as normal peak picking algorithm. In this method, the peaks are extracted one by one from the maximum valued peaks to minimum valued peaks and the number of peaks depends on the order of the sinusoidal model. This method may result in picking up spurious

TH-1853_02610201

peaks. So, the peak picking algorithm can be modified to avoid the harmonic peaks resulting from the fundamental frequency or pitch frequency. In pitch dependent feature extraction, the maximum valued peaks are picked up one by one by considering the logic that the frequency separation between consecutive peaks more than the pitch frequency. Fig. 4.11 shows how eleven peaks are selected from the magnitude spectrum of a pre-emphasized speech signal. By using some minimum frequency difference between successive peaks, the spurious ones can be eliminated. Then the picked peaks are arranged in frequency track and used as a feature vector for speaker recognition.

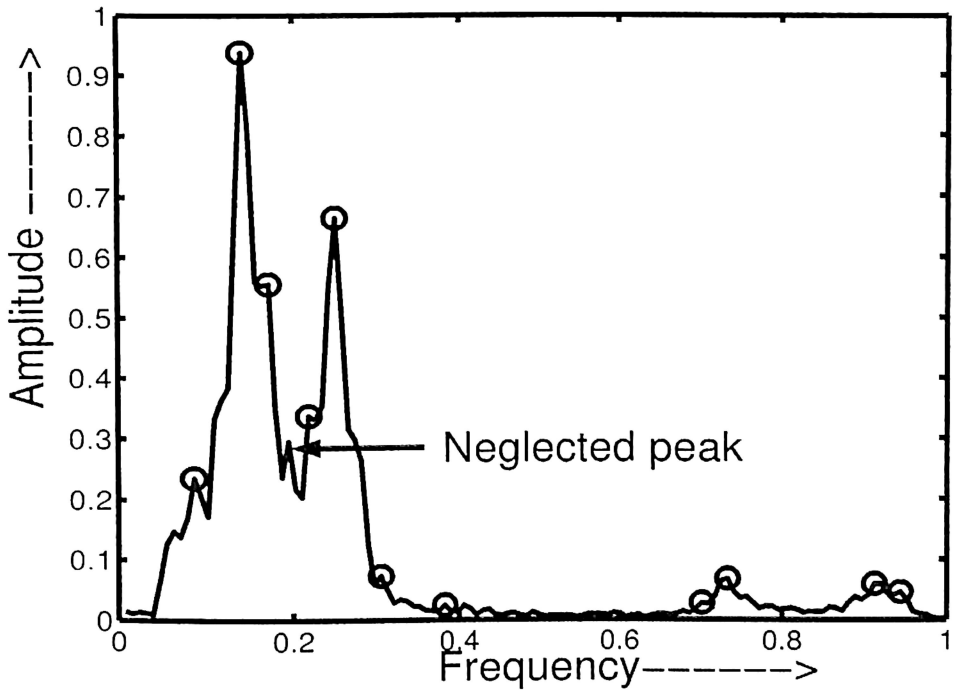


Figure 4.11: Significant peak selection from magnitude spectrum

Fig. 4.12 shows comparison of speaker identification results between normal and pre-emphasized speech using sinusoidal amplitude features (SAF). Speech utterances of twenty speakers from YOHO database are used for this experiment. A 256-size VQ codebook is used for testing the speaker identification rate. Plots in Fig. 4.12 show speaker identification rate versus model order. The identification rate of pre-emphasized speech signal is more than the normal speech signal when the model order is above fourteen. Up to fourteenth order, normal speech or non pre-emphasized speech has higher speaker identification rate. This may be due to the fact speaker specific information is better represented by higher order sinusoidal model for pre-emphasized speech.

The speaker recognition results depend not only on the number of the speakers but also on the

TH-1853_02610201

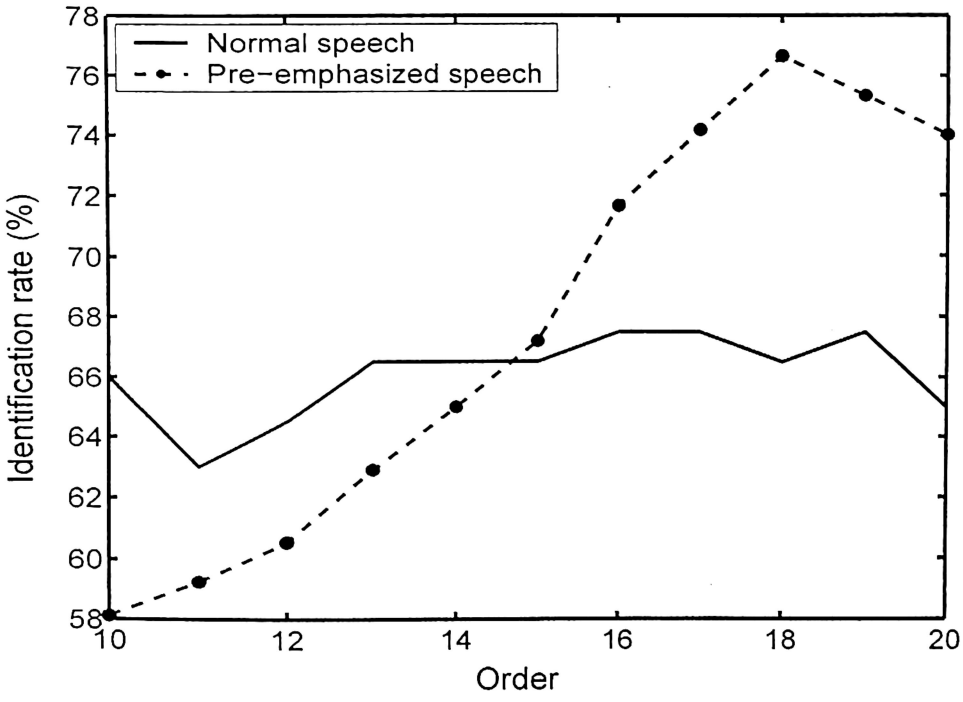


Figure 4.12: Speaker Identification rate versus order

database. In sinusoidal model features, the VQ based speaker identification results are evaluated for speech utterances from YOHO database. Total number of speakers tested is 106. A 256-size VQ classifier is used for evaluation of speaker identification results. Twenty coefficients per feature vector are estimated for each feature. Sinusoidal frequency feature (SFF) shows less performance compared to sinusoidal amplitude feature (SAF). The frequency feature shows 44% correct identification. Correct identification rate for the amplitude feature is 86.42%. For a 12^{th}-order sinusoidal model, the amplitude feature shows a correct identification rate of 85.38%. This shows that there is only two percent increase in speaker identification rate for increase of the model order from ten to twenty. This may be due to the fact that more number low level peaks, which do not carry significant speaker information, are extracted for the higher order model. For comparison with other features, twelve order sinusoidal model is used for the rest of the work.

4.3.1 Independent information analysis

In sinusoidal model, amplitude and frequency feature information are valuable for speaker recognition. These two features contain spectral peak information. The speaker information in these two features may be same or they may contain independent information. Speaker recognition can be improved by

TH-1853_02610201

exploiting any independent or different speaker information present in these two features. In this work, we propose evaluation of confusion probability from confusion matrix for analysis of independent information with SAF and SFF features. Confusion matrices are evaluated for SAF and SFF and they are represented as C_{SAF} and C_{SFF}. Two confusion matrices shown below are evaluated using 5 speaker models and 20 test utterances per speaker. A 256-size VQ codebook is used to represent each speaker model. Each column represents one speaker model and each row represent speaker identification results using 20 test utterances from each speaker. The values in the first row of C_{SAF} shows the speaker identification results when 20 test utterances of first speaker are used. Out of the 20 test utterances, 11 utterances are successfully identified with the first speaker model, 3 utterances are identified with the second speaker model, 3 utterances are identified with the forth speaker model and 3 utterances are identified with the fifth speaker model. As expected the diagonal values are higher compared to the other entries in both confusion matrices. Addition of diagonal values show speaker identification (SI) success rate of the feature. For this experiment, SAF and SFF have SI rates of 66% and 63% respectively. Off-diagonal values in the confusion matrix show how a particular speaker's test utterance is wrongly matched with another speaker model. If both the features have same information, the wrong matching would have relative similar values in both the confusion matrices. It is observed that the relative values in the off-diagonal elements are not same. First element of third row of C_{SAF} has a value of 6 where as its value in C_{SFF} is 0. This shows that third speaker's test data is relatively more matched with the first speaker in case of SAF compared to SFF. Similar results are observed with the other speakers' test data. This proves that the speaker information in SAF is different from the speaker information in SFF. In C_{SAF}, C_{SFF}, P_{SAF}, P_{SFF}, the columns represent the speaker models and the rows represent the speaker of test utterances.

$$C_{SAF} = \begin{pmatrix} 11 & 3 & 0 & 3 & 3 \\ 2 & 15 & 0 & 2 & 1 \\ 6 & 1 & 8 & 2 & 3 \\ 1 & 0 & 0 & 19 & 0 \\ 7 & 0 & 0 & 0 & 13 \end{pmatrix} \qquad C_{SFF} = \begin{pmatrix} 10 & 1 & 2 & 5 & 2 \\ 0 & 11 & 6 & 1 & 2 \\ 0 & 1 & 16 & 2 & 1 \\ 1 & 0 & 0 & 19 & 0 \\ 5 & 0 & 4 & 4 & 7 \end{pmatrix}$$

It may be inadequate to draw the conclusion on independent information content of the features by observing the absolute values of mismatches of speakers' test data. Absolute values of mismatches will depend on the success rate of the corresponding feature. For a feature with higher success rate the mismatch value may be small. For a better logical inference, we evaluate the probability of confusion. P_{SAF} and P_{SFF} are the probability confusion matrices for both the features. First element in the first

TH-1853_02610201

row of $\mathbf{P_{SAF}}$ has a value of 0.33 which show the probability of mismatch of the first speaker's data with the second speaker model. This is evaluated by estimating the probability of mismatch of the first speaker's data with the second speaker model from the values in $\mathbf{C_{SAF}}$. Similarly, the other probability values in $\mathbf{P_{SAF}}$ are estimated. The diagonal values are not considered in this estimation. Each speaker's mismatch probability with the rest four speakers' model are evaluated.

$$\mathbf{P_{SAF}} = \begin{pmatrix} 0.33 & 0 & 0.33 & 0.33 \\ 0.40 & 0 & 0.40 & 0.20 \\ 0.50 & 0.08 & 0.17 & 0.25 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix} \qquad \mathbf{P_{SFF}} = \begin{pmatrix} 0.10 & 0.20 & 0.50 & 0.20 \\ 0 & 0.67 & 0.11 & 0.22 \\ 0 & 0.25 & 0.50 & 0.25 \\ 1 & 0 & 0 & 0 \\ 0.40 & 0 & 0.30 & 0.30 \end{pmatrix}$$

The difference probability confusion matrix can be found out as

$$\mathbf{P_D} = |\mathbf{P_{SAF}} - \mathbf{P_{SFF}}| \tag{4.3.1}$$

In this work, we propose the addition of the values in $\mathbf{P_D}$ as a measure for the independent information between the features. Probability of independent information between the features is defined as

$$\mathbf{I_{IN}} = \frac{1}{10} \sum_{i=1}^{5} \sum_{j=1}^{4} p_D(i,j) \tag{4.3.2}$$

where $p_D(i,j)$ is the value of the (i,j)-element in $\mathbf{P_D}$. The summation limits are chosen as five and four because, in this experiment, we have five speakers' data tested with four other speaker models. The normalization factor of ten is chosen because the maximum possible value of the summation will be ten in this experiment with five speaker models. The value of $\mathbf{I_{IN}}$ will vary from 0 to 1. A value of 0 will indicate identical information between the features. Higher value of $\mathbf{I_{IN}}$ will indicate higher independent information content between the features. From the above experiment with five speakers, we obtain $\mathbf{I_{IN}} = 0.429$. This value prove that there may be considerable amount of independent information between the sinusoidal amplitude feature (SAF) and sinusoidal frequency feature (SFF). This independent information can be exploited for improvement of speaker recognition performance.

Following test is conducted to show the improvement in speaker verification by combining the Amplitude and the Frequency features. Speaker verification system accuracy is mostly mentioned by equal error rate. The equal error rate of amplitude and frequency features are calculated separately. We have tested the equal error rate for 106 male speakers from YOHO database. The equal error rate of the

TH-1853_02610201

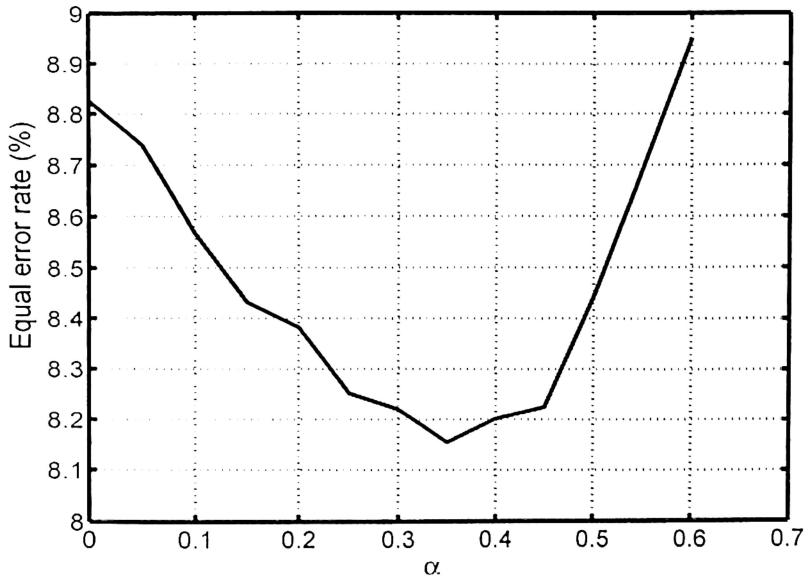


Figure 4.13: Equal error rate of combined confidence score of SAF and SFF

amplitude feature is less than the frequency feature. To exploit the independent information between Amplitude and Frequency features, we combined the confidence scores and evaluated the equal error rate. Each feature's confidence score is calculated by exponential of negative of mean square error (d), given as $CS_{fe} = exp(-\mathrm{d})$, where fe represents the feature name. Two speech feature recognition systems can be combined by linear weighted confidence scores. Combined confidence score can be given by $CS = \alpha CS_{SAF} + (1 - \alpha)CS_{SFF}$. Percentage equal error rate curve has been plotted in Fig. 4.13 for combined SAF and SFF. The combined equal error rate is same as the equal error rate of SFF when the value of α is zero. Similarly, for value of α equal to one, the combined equal error rate is same as the equal error rate of SAF. The value of combined equal error rate is less than the individual equal error rate values of SAF and SFF. Minimum value of equal error rate is obtained for $\alpha = 0.35$. This result shows that some independent information present in both the features help improve the speaker verification performance.

4.3.2 Speaker recognition under stressed condition

In this subsection, we present speaker recognition results with sinusoidal model features. Sinusoidal model features are evaluated for speaker recognition under stressed condition. Pre-emphasized and non pre-emphasized speech are used for these analysis. Table 4.2 shows speaker identification (SI) and speaker verification (SV) results using vector quantization. SAF feature's speaker verification

TH-1853_02610201

results with the pre-emphasized speech is obtained as 20.44%. But its value when the coefficients of the SAF feature is taken as normal amplitude values instead of log magnitude square, the verification result improves and the equal error rate is obtained as 14.36%. Log magnitude square value of the amplitude reduces the dynamic range of the feature which may be the reason for reduction of speaker verification performance. Hence, for speaker verification, we have used amplitude values as the SAF features instead of log magnitude square values. It is observed that SAF feature performs best among the three sinusoidal features for speaker identification and speaker verification with non pre-emphasized and pre-emphasized speech. Sinusoidal phase feature (SPF) shows the lowest recognition performance. Speaker specific vocal articulation information are expected to have correlation with the phase function. Low recognition values obtained with the SPF feature requires further investigation on SPF for its use in speaker recognition problem.

Table 4.2: Speaker recognition results under stressed condition with normal(Neutral emotion) speaker models

SI (Pre-emphasized)	SI (Non pre-emphasized)	SV (Pre-emphasized)	SV (Non pre-emphasized)
SAF(35.31)	SAF(34.52)	SAF(14.36)	SAF(13.37)
SFF(34.12)	SFF(25.39)	SFF(27.50)	SFF(34.95)
SPF(15.07)	SPF(11.90)	SPF(41.79)	SPF(44.20)

Table 4.3: Speaker identification rate (%) for sinusoidal amplitude feature with log magnitude squared coefficients.

Feature: SAF (log magnitude square)					
Training (16 size codebook)	**Testing**				
	Angry	Lombard	Neutral	Question	Average
Angry	52.38	23.8	19.04	17.46	28.17
Lombard	20.63	60.31	31.74	31.74	36.11
Neutral	20.63	26.98	60.31	33.33	35.31
Question	12.69	23.8	44.44	52.38	33.33

Table 4.3 shows speaker identification results for different stressed conditions or emotions. Four codebooks for each speaker are designed and utterances from four emotions are tested with these four codebooks. Speech utterances from SUSAS database are used for evaluation these results. A 16-

TH-1853_02610201

Table 4.4: Speaker identification rate (%) for sinusoidal amplitude feature with amplitude values coefficients.

Feature: SAF (amplitude)					
Training (16 size codebook)	**Testing**				
	Angry	Lombard	Neutral	Question	Average
Angry	41.26	25.39	9.52	12.69	22.22
Lombard	11.11	47.61	25.39	33.33	29.36
Neutral	19.04	25.39	42.85	42.85	32.53
Question	22.22	9.52	42.85	33.33	26.98

Table 4.5: Speaker identification rate (%) for sinusoidal frequency features.

Feature: SFF					
Training (16 size codebook)	**Testing**				
	Angry	Lombard	Neutral	Question	Average
Angry	47.61	22.22	26.98	25.39	30.55
Lombard	22.22	55.55	26.98	17.46	30.55
Neutral	14.28	34.92	57.14	30.15	34.12
Question	25.39	36.5	44.44	38.09	36.11

size VQ classifier is used for evaluation of speaker recognition results. We have chosen the 16-size VQ as this has been used for evaluation speaker recognition results with other features in the previous chapter. SAF feature coefficients are taken as log magnitude square values of the spectral peak values. The results show 60.31% correct identification rate with the Lombard and Neutral speech utterances. Maximum average identification rate of 36.11% is obtained with the Lombard codebook. For speaker verification, normal amplitude values are chosen as SAF coefficients. If normal amplitude values are chosen as SAF coefficients, the speaker identification results are expected to be different from the results shown in Table 4.3. Table 4.4 shows speaker identification results with sinusoidal amplitude feature where the coefficients are chosen as normal amplitude values. The speaker identification performance is poor compared to the previous case. Except for test utterances from Question emotion, results show maximum recognition when the test utterances are from the same emotion. Question emotion has maximum recognition with the Neutral codebook. Lombard speech shows maximum recognition of 47.61%. This may be due to the fact that the voiced part of the speech is more stressed in case of Lombard speech. Neutral codebook shows maximum average recognition of 32.53%. Table 4.5 shows speaker identification results with the sinusoidal frequency features (SFF). SFF shows maximum recognition when the test utterances are from the same emotion. For the frequency feature, Neutral emotion shows maximum speaker identification result of 57.14%. Question codebook shows maximum average

TH-1853_02610201

recognition of 36.11%.

Speaker identification and verification results of MFCC, LP, LPCC, LAR, ARC and RC are discussed in the previous chapter. Same results are shown with sinusoidal model features in Table 4.6 for comparison of relative performance of nine features. Pre-emphasized speech is used for this experiment. A 16-size VQ classifier is used for evaluation of speaker recognition results. For GMM classifier, 16 Gaussian mixtures are used. For all the features, twelfth order model is used for comparison of their performances. Average speaker recognition results for all the four stressed or emotion test data with Neutral codebook are shown in the table. In case speaker identification, with both the VQ and GMM classifiers, sinusoidal model features perform lower compared to the other features. In case of speaker verification, sinusoidal amplitude feature (SAF) perform as the second best feature. MFCC feature shows least error rate and followed by SAF.

Table 4.6: Feature analysis for speaker recognition under stressed condition (Pre-emphasized)

Rank	VQ based Speaker identification rate(%)	VQ based Speaker verification (EER in %)	GMM based speaker identification rate(%)	GMM based speaker verification (EER in %)
1	ARC (54.75)	MFCC(14.06)	RC(55.15)	MFCC (9.7)
2	MFCC (53.96)	SAF(14.36)	ARC(54.75)	SAF(12.5)
3	RC (53.96)	RC(16.84)	MFCC(53.56)	RC(13.39)
4	LAR (50.79)	ARC(16.47)	LPCC(53.17)	ARC(13.79)
5	LPCC (48.01)	LAR(19.20)	LAR(51.18)	LAR (14.83)
6	LP (43.64)	LPCC(21.58)	LP(49.59)	LPCC(16.82)
7	SFF(34.12)	SFF(27.50)	SFF(38.09)	LP(21.78)
8	SAF(32.53)	LP(29.54)	SAF(32.14)	SPF(35.49)
9	SPF(15.07)	SPF(41.79)	SPF(9.1)	SFF(40.75)

TH-1853_02610201

4.4 Combination of MFCC and Sinusoidal Features

The sinusoidal model features like Amplitude and Frequency features are examined for speaker recognition. The performance of the sinusoidal amplitude feature (SAF) is better than the frequency feature (SFF). But, the Amplitude feature performance is comparatively poor with MFCC. MFCC features are estimated using Fourier transform of mel-frequency band energies in the magnitude spectrum. The peak information in the spectrum is not directly used in case of MFCC feature as the spectral energy over a frequency band is used. Spectral peaks have speaker information. A combination of MFCC and sinusoidal features (SAF and SFF) may provide better speaker recognition performance. In this section, we have tested the speaker verification by combination of MFCC and sinusoidal features in the similar manner as discussed for combination of SAF and SFF in subsection 1.3.1. However, it would be emphasized at this point that still SAF and MFCC are closely related due to their similar origin. The difference between these two is SAF has peak information and MFCC is obtained by using discrete cosine transform (DCT) on the filter bank energies. Each feature's confidence score is calculated by exponential of negative of mean square error (d), given as $CS_{fe} = exp(-\mathrm{d})$, where fe represents the feature name. Two speech feature recognition systems can be combined by linear weighted confidence scores. Combined confidence score of MFCC and SAF can be given by $CS = \alpha CS_{MFCC} + (1-\alpha)CS_{SAF}$. Similarly, MFCC and SFF can be combined by replacing C_{SAF} with C_{SFF} in the expression for CS. Speaker verification results with 106 male speakers' speech utterances from YOHO database are used for evaluation of verification results with the combination of features. Fig.4.14 shows variation of percentage equal error rate plots for three different combinations of the MFCC, SAF and SFF. It is observed that for combination of MFCC and SAF, the equal error rate is least at α equal to 0.37. The equal error rate reduced from 0.55% to 0.36% when α value increases from 0 to 0.37. But the miss probability at false alarm is equal to 0.01% increases from 4.1% to 5%. So, all these values like equal error rate, miss probability and False alarm probability are considered for the selection of α value. This reduction of percentage equal error rate which shows improvement in speaker verification proves that the sinusoidal amplitude feature (SAF) contains some speaker information which are not captured by MFCC feature. Examination of combination of MFCC with sinusoidal frequency feature (SFF) shows that the equal error rate is least at α equal to 0.53. The equal error rate reduced from 0.55% to 0.39% when α value increases from 0 to 0.53. This improvement in speaker verification shows that there are some independent information between the MFCC and SFF. Comparison of the combination of SAF and MFCC and the combination of SFF and MFCC show slightly better result for combination of SAF

TH-1853_02610201

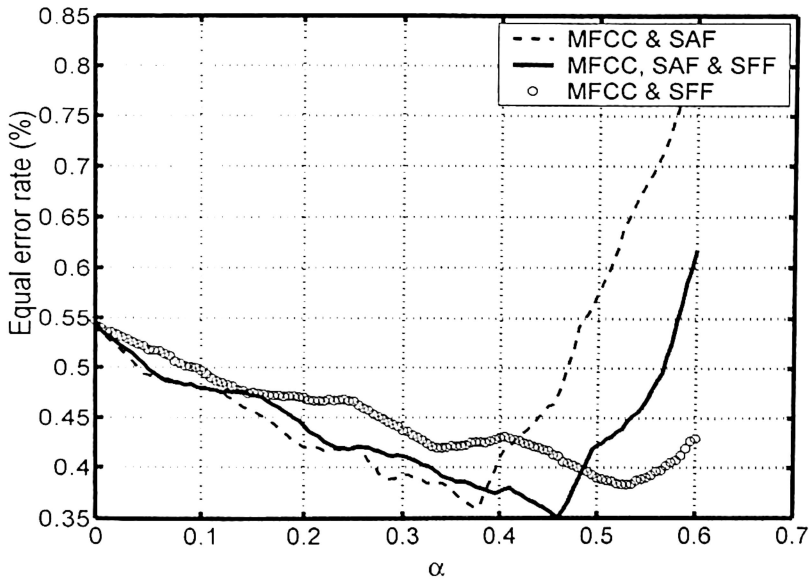


Figure 4.14: Equal error rate of combined confidence score for different combination.

and MFCC. The third curve in Fig.4.14 shows results for combination of MFCC with SAF and SFF. For this experiment, first SAF and SFF are suitably combined by finding out the minimum equal error rate. Then this combination of these sinusoidal features is combined with the MFCC feature. Combination of sinusoidal amplitude feature (SAF) and sinusoidal frequency feature (SFF) results are shown in Fig. 4.13 and discussed in section 1.3.1. The equal error rate decreases from 8.81 to 8.15. as α value increases from 0 to 0.35. When sinusoidal features are combined with the MFCC feature, the equal error rate reaches a minimum value of 0.35 for an α value of 0.46. This result does not show much improvement from the combination of MFCC and SAF. This may be due to the fact that the quantity of extra information may be same in SAF and SFF compared to MFCC.

Above idea of combining MFCC with SAF can be extended to speaker recognition under stressed condition. Fig. 4.15 shows variation of percentage equal error rate for speaker verification test under stressed condition with different combinations of MFCC and SAF. It is observed that minimum equal error rate of 12.52% is obtained. This value is less than the equal error rate values of individual features. Also, this value is more than the minimum equal error rate obtained with the speech utterances in YOHO database which is discussed in the previous paragraph. For stressed speech signal the speaker verification results can be improved by suitably combining the MFCC and sinusoidal features. This idea has been experimented in the next chapter for improvement of speaker identification under stressed condition.

TH-1853_02610201

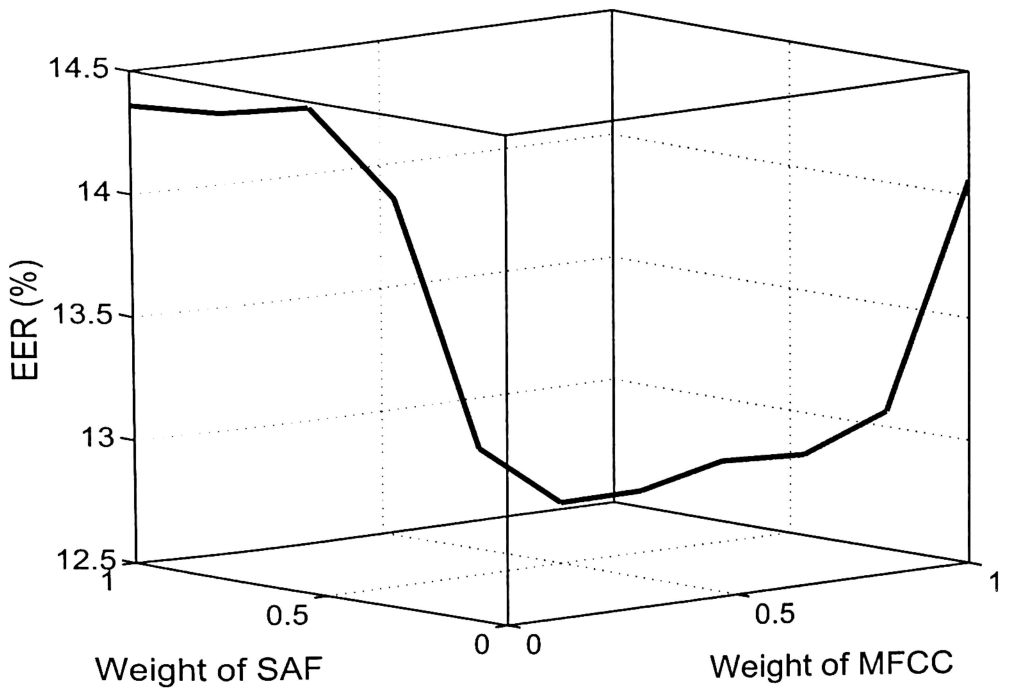


Figure 4.15: Equal error rate of MFCC and SAF for stressed speech.

TH-1853_02610201

Chapter 5

Stress Compensation for Speaker Recognition

The characteristics of the speech signal are modified due to the stressed condition or emotion. Hence, the performance of speaker recognition system is effected by the stressed condition or emotion. In most of the cases, it results in degradation of performance of the system. In chapter 3, six speech features are analyzed for evaluation of their relative performance in speaker recognition for speech signals under four stressed conditions. Sinusoidal model based features are tested for speaker recognition under stressed conditions in chapter 4. Analysis of stressed speech shows that the degradation of speaker recognition system can be compensated. Compensation techniques are suggested for speech recognition under stressed conditions. They are feature or parameter based compensation techniques [127], [84] and prosody modification [91], [128] based compensation techniques. The parameter based compensation technique converts the stressed condition into normal condition in feature domain. In prosody modification, the pitch and duration can be changed according to the stress variations. Modification of pitch and duration transforms the stressed speech signal into normal speech signal.

In this chapter, four compensation techniques are proposed and tested for speaker recognition under stressed condition. In chapter 3 and 4, it is observed that MFCC feature is the best feature for speaker recognition under stressed condition. In this chapter, the proposed compensation techniques are tested with MFCC features. Figure 5.1 shows the block diagram of a speaker recognition system with stress compensation technique. The goal of the compensation technique block is to reduce the stress effect present in the speech signal. Four compensation techniques are proposed and tested for improvement of speaker recognition under stressed condition. In Section 5.1, we propose speaker and stress information based compensation (SSIC) technique. Section 5.2 presents compensation by removal of stressed vector (CRSV). Section 5.3 presents combination of MFCC and SAF (CMSA) as a compensator. Cepstral mean normalization (CMN) is used for compensation in Section 5.4. Finally, we present selection of

TH-1853_02610201

compensator by stress recognition in Section 5.5.

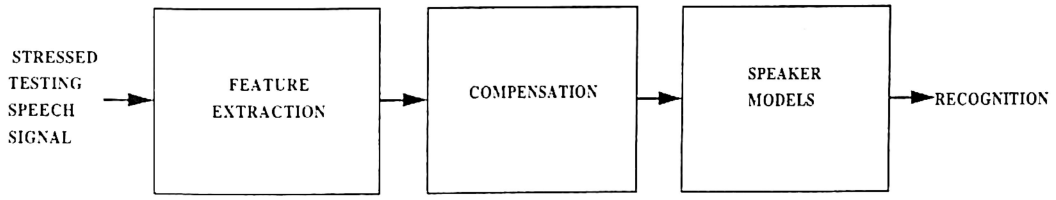


Figure 5.1: Speaker recognition system with a stress compensation technique

5.1 Speaker and Stress Information based Compensation(SSIC)

F-ratio values of four stressed conditions of nine speakers are analyzed in Chapter 3. Higher F-ratio value of a coefficient between the speakers indicate good speaker recognition performance. Higher F-ratio value of a coefficient between the stressed conditions shows the higher potential of the coefficient as a carrier of stressed information. Figure 3.6 shows F-ratio values between speakers and between stressed conditions from the same speech data. It is observed that the F-ratio values between the speakers are different from the F-ratio values between the stressed conditions. An important observation is that for most of the features the relative F-ratio values between the coefficients show an inverse property. This means coefficients, which have significant speaker information may not be relatively important from the point of view of stress information. For example, F-ratio value of eighth coefficient of MFCC feature has a relatively higher value compared to other coefficients when evaluated between the stressed conditions. Same eighth coefficient of MFCC feature has a relatively lower F-ratio value compared to other coefficients when evaluated between the speakers. This result suggest that this coefficient is most effected by the stress conditions or it is a good carrier of information related to stressed conditions. Similarly, it can be concluded that it is a poor carrier of speaker information.

Suppression or removal of a coefficient which shows strong stress content may improve the speaker recognition performance because it will help eliminate the variations due to stress condition in a speaker's speech data. Throughout the work, we have considered 12 coefficients. Only in compensation analysis 14 coefficients are considered. This is with the objective of removing about two coefficients which are affected significantly from stress. However, for reference and uniformity, results of fourteen coefficients case are shown for 12 coefficients given in Table 5.4. F-ratio values of fourteen MFCC coefficients are shown in Table 5.1. F-ratio values shown in the second column are evaluated between four emotions or stressed conditions. F-ratio values shown in the third column are evaluated between

TH-1853_02610201

nine speakers using the same speech data. Higher F-ratio value indicates the relatively higher potential of the coefficient for stress or speaker information if the F-ratio values are evaluated between the stress conditions or between the speakers respectively. From the tabulated F-ratio values, it can be concluded that the relative importance of the coefficients is different in case of stress from that in case of speaker. Fourth column in Table 5.1 shows the difference of F-ratio values between the stress and speaker. Tenth coefficient has maximum difference value followed by the eighth coefficient. This result indicates that the tenth coefficient may be most responsible candidate for degradation of speaker recognition due to stress condition present in the signal. Suppression of this coefficient may improve speaker recognition results. The SI results with fourteen coefficients are shown in the first row of Table 5.2. The reference codebook is a 16-size VQ based Neutral codebook. Hence, the Neutral test speech has maximum recognition of 84.12%. Second row shows SI results with 13 MFCC coefficients. These 13 coefficients are chosen from the original 14 coefficients after removal of the lowest F-ratio valued coefficient of the speakers. From third column of Table 5.1, 13^{th} coefficient has the lowest F-ratio value. The results show that speaker identification is improved in case of Angry emotion. For Lombard emotion, the results remain same as in case of uncompensated case. For Neutral and Question emotion the SI results degrade. The average SI result with 13 coefficients is lower compared to the uncompensated results with 14 coefficients. Next, 12 coefficients are chosen for testing SI value by removing the 13^{th} and 10^{th} coefficients. It is observed that the SI results do not improve compared to the uncompensated case. Above results prove that removal of coefficients based on their relative values may not help improve speaker recognition under stressed condition. Third row in Table 5.2 shows SI results after removal of first one coefficient and then two coefficients. In this case, the coefficients are removed based on F-ratio difference shown in third column of Table 5.1. Tenth coefficient has a maximum difference value of 0.0154 and eighth coefficient has next difference value of 0.009. The average SI value has improved to 53.17% with 12 coefficients. For Angry emotion, the speaker identification (SI) result has gone up to 31.74%, after removal of tenth and eighth coefficients, from 20.63% in case of uncompensated case.

Due to removal of coefficients, some speaker information are also removed from the feature vectors. As a result, though there is improvement in case of Angry speech, SI results degrade in case of Neutral and other emotions. This requires further investigation or modification of the above compensation technique. Fig. 5.2 shows plot of F-ratio values of MFCC feature evaluated between the speakers. These F-ratio values are evaluated from the stressed speech data. It is observed that higher order coefficients have relatively less F-ratio values compared to lower order MFCC coefficients. Hence, higher order coefficients have less speaker information compared to lower order coefficients. This knowledge can be

TH-1853_02610201

Table 5.1: F-ratio of fourteen MFCC coefficients(%)

Coefficient number	Stress (A)	Speaker (B)	A-B
1	0.0702	0.0998	-0.0296
2	0.0344	0.0572	-0.0228
3	0.0297	0.0634	-0.0337
4	0.0073	0.0689	-0.0616
5	0.0002	0.0783	-0.0781
6	0.0073	0.0237	-0.0164
7	0.0197	0.0361	-0.0164
8	0.0376	0.0286	0.009
9	0.0231	0.0488	-0.0257
10	0.023	0.0076	0.0154
11	0.0099	0.0091	0.0008
12	0.003	0.0429	-0.0399
13	0.0043	0.0066	-0.0023
14	0	0.0327	-0.0327

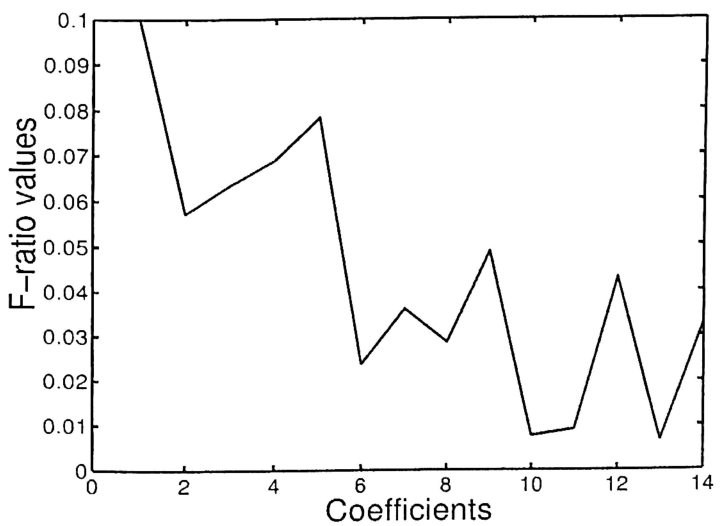


Figure 5.2: F-ratio values of MFCC feature

used for improvement in speaker recognition in the following manner. A compensator implemented at the decision stage of the codebook by evaluating a modified error distance. The modified error distance is implemented by dividing the codebook errors with the F-ratio values. For example, if the testing vec-

TH-1853_02610201

Table 5.2: Speaker identification rate (%) with F-ratio test analysis

Coefficients	Testing				
	Angry	Lombard	Neutral	Question	Average
14c	20.63	47.61	84.12	58.73	52.77
13c (speaker based coefficient removal)	25.39	47.61	79.36	57.14	52.38
12c (speaker based coefficient removal)	19.04	44.44	80.95	53.96	49.60
13c (stress based coefficient removal)	23.80	44.44	82.53	57.14	51.98
12c (stress based coefficient removal)	31.74	44.44	79.36	57.14	53.17

tor and codebook vector are $\vec{x} = [a_1, a_2...a_{14}]$ and $\vec{c} = [c_1, c_2...c_{14}]$ respectively, the distance between codebook vector and testing utterance vector ($(\vec{x} - \vec{c})^2$) are weighted by inverse F-ratio value of each coefficient. Hence the modified error distances are ($\frac{1}{f_1} \times (a_1 - c_1)^2$, $\frac{1}{f_2} \times (a_2 - c_2)^2 ... \frac{1}{f_{14}} \times (a_{14} - c_{14})^2$). Here, f_i represents F-ratio value of the i^{th} coefficient. This results in higher multiplication factor for the lower F-ratio valued coefficients. In this way the error distances for the lower valued coefficients are enhanced. Second row of Table 5.3 shows the speaker identification rate for this modified method of evaluation of error distances. Reference codebook is a 16-size VQ based Neutral speaker codebook. It is observed that the speaker identification result for Angry and Question emotion has increased. The average speaker identification result has improved to 57.34% from 52.77% in case of uncompensated case.

Fig. 5.3 shows the multiplication factors of the error distance for corresponding coefficients. X-axis in Fig. 5.3 shows the coefficient number and the corresponding values in this figure show the multiplication factors. Lower F-ratio valued coefficients have higher multiplication factors. Most of the lower valued F-ratio coefficients are higher order MFCC coefficients. Higher order MFCC coefficients are associated with source information [80]. Source is relatively more responsible for the emotion present in the speech signal compared to the vocal tract [130]. In the above experiment only speaker discrimination information is used for finding the multiplication factors of the coefficients. Stress information is present with the F-ratio values evaluated between the stress conditions or emotions. Speaker recognition under stressed conditions can be improved if the stress information present with the F-ratio values of stress are properly exploited or used. For further experimentation, the multiplication factors associated with the lower F-ratio valued coefficients are reduced and shown as m2 and m3 characteristics in Fig. 5.3. For generating m2 characteristic, 0.5 times the mean of F-ratio values evaluated between the four stress conditions is added to the F-ratio values of the speakers and then

TH-1853_02610201

the reciprocal of these values are taken as multiplication factors. Similarly, for m3 characteristic, 0.2 times the mean of F-ratio values of stress is added to the F-ratio values of speakers before taking the reciprocal of their values as multiplication factors. Speaker identification (SI) results for m2 and m3 multiplication factors are shown in third and fourth rows of Table 5.3. It is observed that the average speaker identification rate has improved compared to the uncompensated case. For m2, its value is 57.14% and for m3, its value is 58.32%. With m2 multiplication factor SI results improved for Angry and Lombard emotions. Further improvement in Angry, Lombard and Question emotion is observed with the m3 multiplication factor. The improvements in SI results prove that best results can be obtained by suitably using both the speaker and stress information present in the signal.

Above results are obtained with each MFCC feature vector having 14 coefficients. It is observed that speaker identification improves if both speaker and stress information are used. Stress information is used by adding some percentage of average F-ratio values for stress with F-ratio values of speakers for choosing two suitable multiplication factors, m1 and m2. Same experiment is carried out with 12 coefficients for each MFCC feature vector. For selecting a suitable multiplication factor the percentage of average F-ratio value for stress is varied and the corresponding average speaker identification results are evaluated. Fig. 5.4 shows average SI results with different stress weights. Different stress weights

TH-1853_02610201

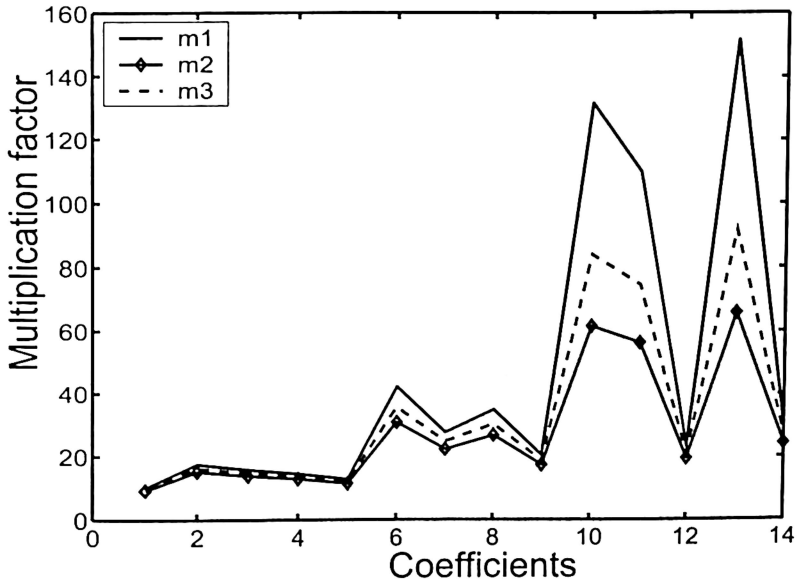


Figure 5.4: Multiplication factors of error distances for corresponding coefficients

Table 5.3: Speaker identification rate (%) for different multiplication factors with the error distances in the codebook (14 coefficients)

Compensators / Multiplication factors	Testing				
	Angry	Lombard	Neutral	Question	Average
Un compensated	20.63	47.61	84.12	58.73	52.77
m1	36.50	47.61	80.95	60.31	56.34
m2	31.74	53.96	82.53	60.31	57.13
m3	34.92	57.14	82.53	60.31	58.73

shown in X-axis of Fig. 5.5 are the values with which the average F-ratio value is normalized before adding it to the F-ratio values of the speakers. Average SI results vary with different stress weights. It is observed that maximum average SI value of 56.34% is obtained with a stress weight value of 6. This value of stress weight is chosen for m2 multiplication factor in case of 12 MFCC coefficients. Fig. 5.6 shows two multiplication factors for the error distances. The multiplication factor, m1, is taken as reciprocal of F-ratio values of speakers. No stress factor is considered for m1 multiplication factor. Addition of stress information to the speaker information reduces the dynamic range of the multiplication factor which is evident from the characteristic of m2. This help improve the SI results. Table 5.4 shows SI results with 12 MFCC coefficients. It is observed that the average SI results for m2

621.3
RAJ/F
P07
THESIS
Lakshminath Bezbaroa Central Library
Indian Institute of Technology Guwahati
ACC. No. TH. 1853
Date 28/3/19
TH-1853_02610201

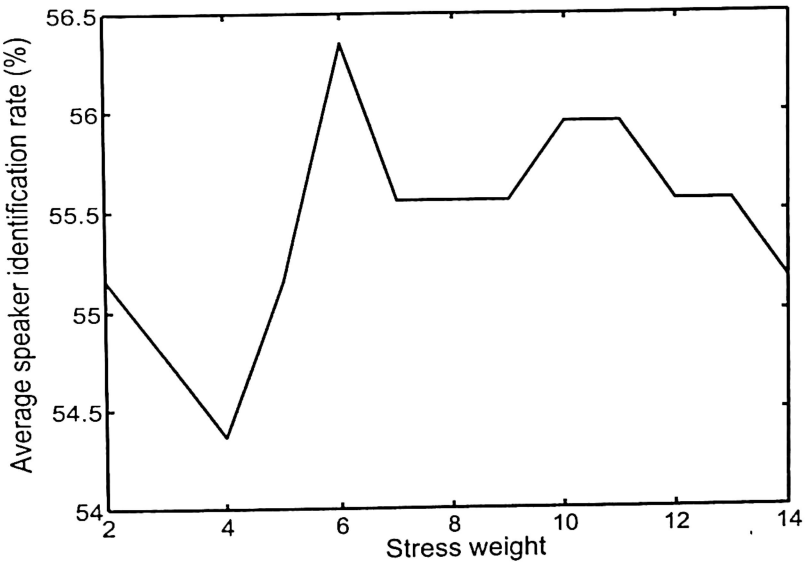


Figure 5.5: Average speaker identification with different values of stress weights

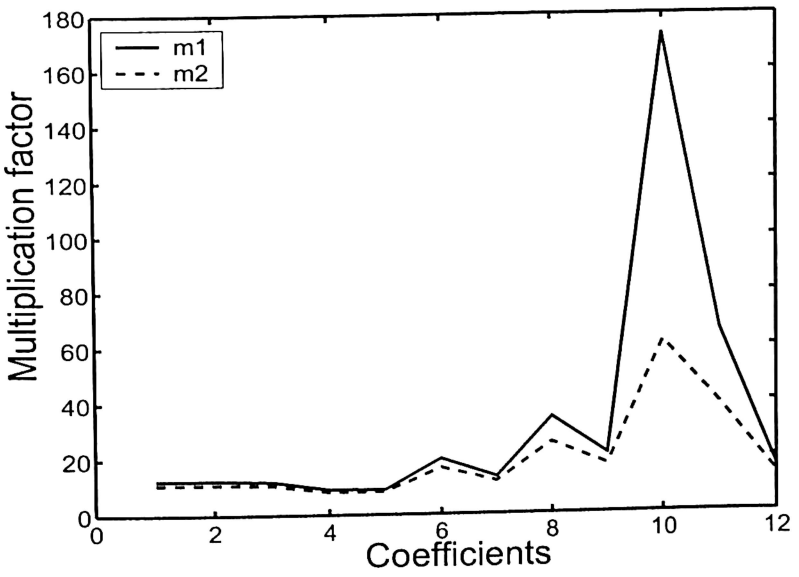


Figure 5.6: Multiplication factors of error distances

multiplication factor is higher compared to the SI results with uncompensated and m1 multiplication factor. Best SI results of 30.15%, 50.79% and 84.12% are obtained for Angry, Lombard and Neutral emotion speech data respectively. Best SI result of 61.90% for Question emotion is obtained with m1 multiplication factor. These results prove that lower F-ratio valued coefficients play a major role compared to the higher F-ratio valued coefficients for stress conditions in a speech signal. As lower F-ratio valued coefficients appear for higher MFCC coefficients, this supports the idea which suggests

TH-1853_02610201

that the emotion in a speech signal is mostly controlled by the source [130].

Table 5.4: Speaker identification rate (%) for different multiplication factors with the error distances in the codebook (12 coefficients)

Compensators / Multiplication factors	Testing				
	Angry	Lombard	Neutral	Question	Average
Un compensated	26.98	49.2	80.95	58.73	53.96
m1($\frac{1}{F-ratio_{speakers}}$)	26.98	50.79	77.77	61.90	54.36
m2($\frac{1}{F-ratio_{speakers}+0.5*mean(F-ratio_{stress})}$)	30.15	50.79	84.12	60.31	56.34

5.2 Compensation by Removal of Stressed Vectors (CRSV)

Speaker recognition can be improved if the stress effect is reduced. Stress effect on the features can be estimated by evaluating the distance between the feature vectors of the stressed speech and the Neutral speech. Feature vectors effected by stress are expected to be at larger distances compared to the feature vectors which are not effected by the stress condition. In this compensation technique, stressed feature vectors are detected by comparing the distance with a predefined threshold. In the proposed compensation technique, Common Neutral Codebook (CNC) is developed using the training feature vectors of Neutral speech data from all the speakers. The CNC vectors contain average Neutral speech information. The speaker specific information is suppressed in CNC vectors. Test vectors, ($X = [\vec{x}_1, \vec{x}_2 ... \vec{x}_T]$), which are close to CNC vectors, ($C = [\vec{c}_1, \vec{c}_2, ... \vec{c}_P]$), can be considered as vectors without much speaker specific information. This compensation technique is shown in Fig. 5.6. In this compensation technique, 16 size CNC has been prepared for all the speakers. Feature vectors ($\vec{x}_i$), from speech utterances of three emotions, Angry, Lombard and Question, are compared with the CNC vectors. The minimum distance, ($D_i = \min_{\vec{c}_j \epsilon C} d^2(\vec{x}_i, \vec{c}_j)$), is evaluated between the emotion vector $\vec{x}_i$ and the CNC vector, $\vec{c}_j$. The minimum value of all these distances (D_i) is determined and it is considered as a threshold for determining the stressed vectors. During testing, each test vector is compared with the CNC vector. If the distance between these vectors is more than the threshold the tested vector is retained. Those vectors for which the distance is less than the threshold value, they are dropped from further processing. Compensation by removal of stressed vectors is proposed using the above logic. Neutral speaker model is trained using feature vectors from Neutral training utterances.

TH-1853_02610201

In this experiment minimum distance as threshold has not shown any improvement as this includes all the test vectors. We have experimented with different thresholds which are taken after multiplying the minimum distance with some multiplication factor greater than one. Best results are obtained with a multiplication factor of 16. So, the threshold is set as sixteen times the threshold of the minimum distance. The results are shown in Table 5.5. From the table it can be observed that the performance improves only in the Lombard speech case from 49.2% to 52.38%. As a result, the overall improvement of average recognition performance increases from 53.96% to 54.96%. This technique requires further investigation for use in case of speaker recognition under stressed condition.

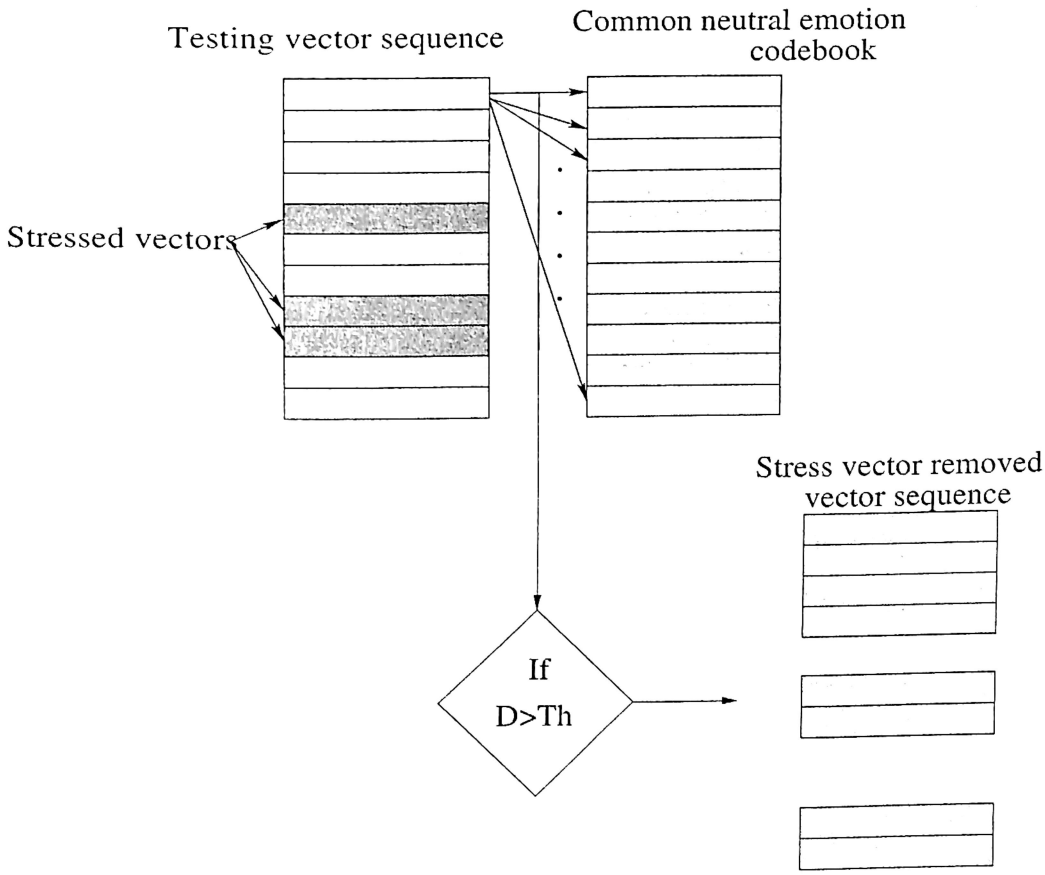


Figure 5.6: Compensation by removal of stressed vector

Table 5.5: Speaker identification rate (%)

Training (16 size codebook)	**Testing**				Average
	Angry	Lombard	Neutral	Question	
Un compensated	26.98	49.2	80.95	58.73	53.96
Stressed vector removal	26.98	52.38	80.95	58.73	54.76

TH-1853_02610201

5.3 Combination of MFCC and Sinusoidal Amplitude (CMSA) Features

In Section 4.4, we showed that the combination of MFCC and sinusoidal amplitude features (SAF) improve the speaker recognition performance under neutral and stressed condition. The experiment is conducted using 106 male speakers of YOHO database. MFCC features are estimated from the magnitude spectrum of the speech signal. The MFCC feature extraction technique basically includes windowing the signal, applying the DFT, taking the log of the magnitude and then warping the frequencies on a mel scale followed by applying the inverse DCT.

In case of sinusoidal amplitude features, the speaker specific features are extracted from the spectral peaks of DFT magnitude spectrum. The peak information in the spectrum is not directly used in case of MFCC features as the spectral energy over the frequency band is used. However, the spectral peaks also contain speaker specific information. This is because the spectral peaks represent formants, pitch and harmonics. These features contain unique speaker dependent characteristics. Hence, combining MFCC with sinusoidal amplitude features may improve the performance. It is shown in [85], sinusoidal model based features are sensitive to emotion. Therefore, the combining MFCC with SAF can be extended to speaker recognition under stressed condition with a suitable compensation technique. One such approach may be identifying the speaker specific features that are sensitive to emotion (Sinusoidal model based features) and remove them before modeling. This is exploited in the present work.

Table 5.6: Maximum speaker identification results (%) for combined MFCC and SAF features

Compensator	**Emotion of testing utterance**				
	Angry	Lombard	Neutral	Question	Average
Un compensated	26.98	49.2	80.95	58.73	53.96
Combined MFCC and SAF (CMSA)	30.15 $\alpha = 0.15$	50.79 $\alpha = 0.65$	85.71 $\alpha = 0.6$	60.31 $\alpha = 0.1$	56.74

Fig. 5.7 shows block diagram of stress compensation for speaker identification using MFCC and SAF features. Each feature's confidence score is calculated by exponential of negative of mean square error (d), given by $CS_{fe} = exp(-\mathrm{d})$, where fe represents the feature name. Two speech feature

TH-1853_02610201

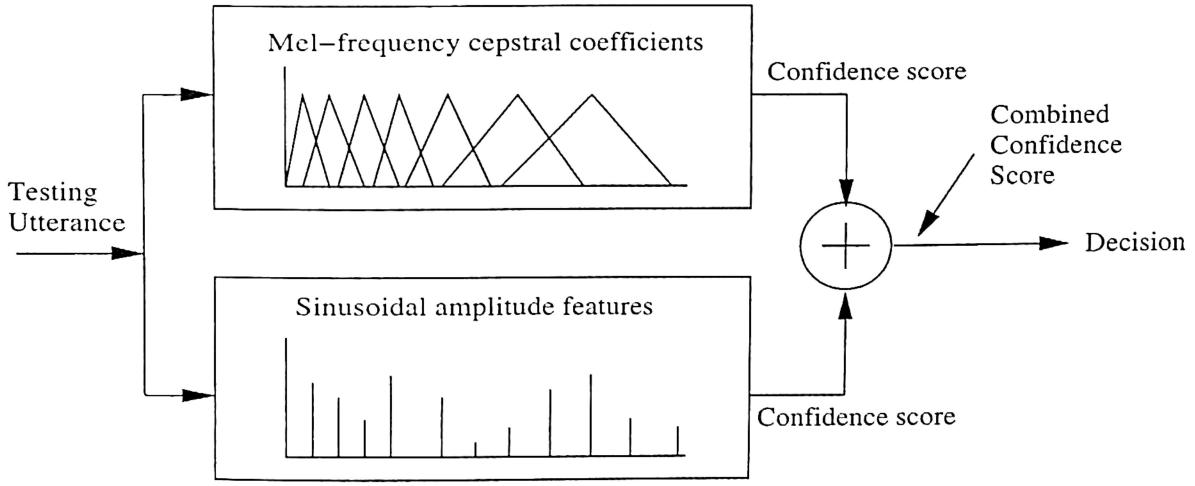


Figure 5.7: Stress compensation using combination of MFCC and SAF features

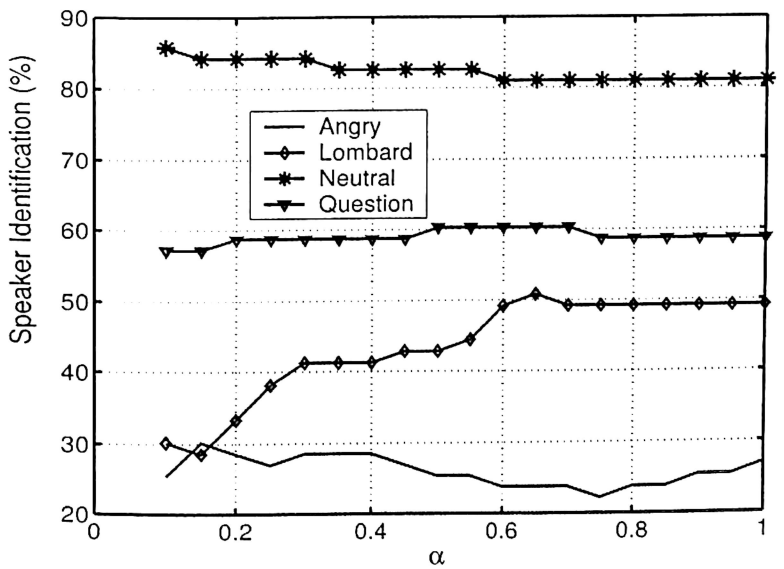


Figure 5.8: Speaker identification results of four emotions

recognition systems can be combined by linear weighted confidence scores. Combined confidence score of MFCC and SAF can be given by $CS = \alpha CS_{MFCC} + (1 - \alpha)CS_{SAF}$. Fig. 5.8 shows speaker identification results for four stressed or emotion speech signals for different values of α. It can be observed that maximum speaker identification results for different emotions are obtained for different values of α. Table 5.6 shows maximum speaker identification (SI) results for this compensator. For Angry and Neutral speech, maximum SI results of 30.15% and 85.71% are obtained respectively with

TH-1853_02610201

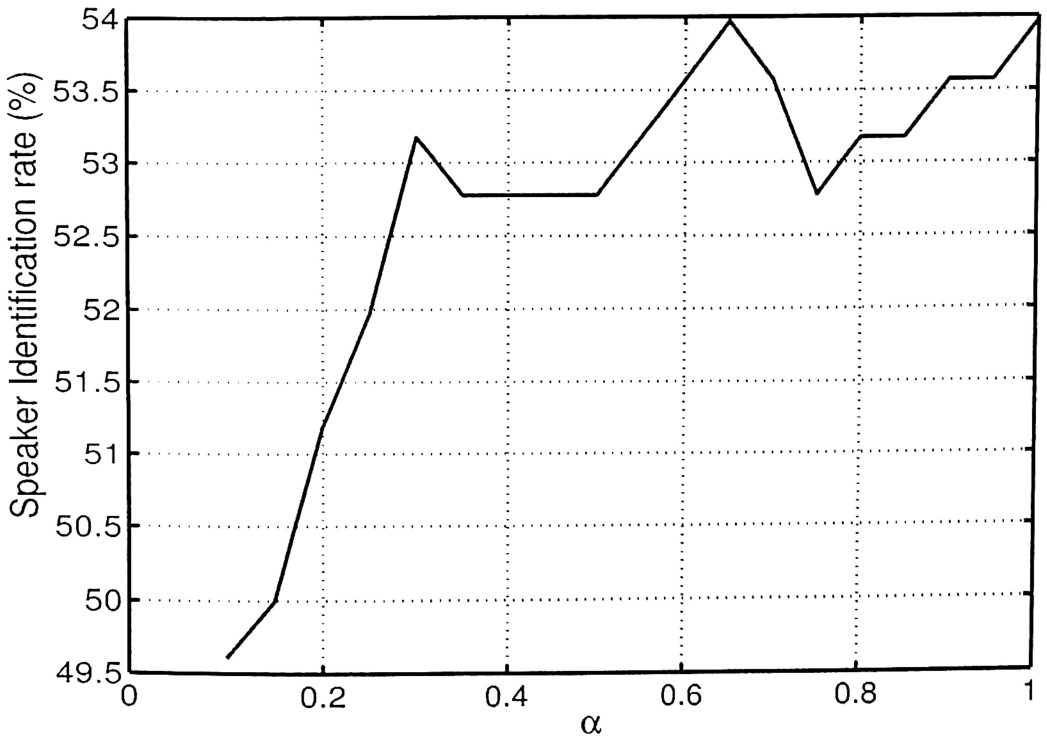


Figure 5.9: Average speaker identification results

α value of 0.1. With a α value of 0.65, maximum SI results of 60.31% and 50.79% are obtained for Question and Lombard emotion speech data. Hence, speaker identification result improves for all the emotions with different values of α. Different values of α can be used if there is prior information about the type of stressed or emotion speech signal. Otherwise, one single α value can be used. This α value can be chosen by finding the maximum average speaker identification. Fig. 5.9 shows variation of average speaker identification rate with different values of α. It is observed that maximum average SI result of 54% is obtained with a α value of 0.65. At this α value Angry, Lombard, Neutral and Question emotions show SI results of 23.81%, 50.79%, 80.95% and 63.31%, respectively. It can also be noted that the average SI result shows an improvement over the uncompensated SI result.

5.4 Cepstral Mean Normalization (CMN)

Cepstral Mean Normalization (CMN) technique is traditionally used to compensate the noise for speech recognition [111]. Speech recognition results are shown to improve by this technique for speech under the stressed condition [94]. In this section, CMN technique is used for possible improvement in speaker recognition under stressed condition. Table 5.7 shows speaker identification results with CMN

TH-1853_02610201

Table 5.7: Speaker identification rate (%)

Normal speaker recognition					
Training (16 size codebook)	**Testing**				
	Angry	Lombard	Neutral	Question	Average
Normal speaker recognition	26.98	49.2	80.95	58.73	53.96
Cepstral Mean Normalization (CMN) (Neutral training utterances)	26.98	41.5	84.84	58.73	52.95
Cepstral Mean Normalization (CMN) (All emotion training utterances)	22.22	46.03	85.71	55.55	52.38

technique. Twelve cepstral coefficients with a sixteen size code book are used for the experiment. The code book is designed with the Neutral speech data. Speech utterances from Angry, Lombard, Neutral and Question emotions are tested with the Neutral speaker code book. First row in Table 5.7 shows results with uncompensated cepstral features. It is observed that Neutral test data shows maximum speaker identification results followed by Question, Lombard and Angry stressed conditions. The average speaker identification with the uncompensated cepstral features is 53.96%. This average SI rate is for all emotion testing utterances with Neutral speaker models. Second row in Table 5.7 shows speaker identification results with CMN technique. In this case the mean value of the cepstral features is estimated from the Neutral training speech data. Cepstral features normalized with a mean value estimated from the Neutral training data is due to the fact that only Neutral training speech data data is assumed to be available before testing. It is observed that the speaker recognition results are not improving as in case of speech recognition [111]. The degradation may be because of less duration utterances. It is observed that the Neutral utterances show improvement from 80.95 % to 84.84 %. The SI rate of Lombard utterances degrade from 49.2 % to 41.5 %. Other two emotion testing utterances maintain same recognition rate. The average SI rate degrades from 53.96 % to 52.95 %. Third row in Table 5.7 show SI results with a CMN technique in which the cepstral mean is estimated from the of all emotions. It is observed that Neutral emotion testing utterances show better performance than mean normalization only by Neutral utterances. But, all other emotions show less recognition than normal recognition performance shown in first row. Hence, it can be concluded that cepstral mean normalization (CMN) technique can improve speaker identification whereas it is not effective in case of stressed speech conditions.

TH-1853_02610201

5.5 Selection of Compensation by Stress Recognition (SCSR)

Four compensators are proposed and tested for improvement of speaker recognition under stressed condition. For evaluation of speaker identification results, 16-size Neutral speaker codebooks with MFCC features are used. It is observed that different SI results are obtained for different stressed conditions or emotions. Table 5.8 shows SI results for 12 order MFCC features. CMSA compensator shows maximum average SI value of 56.74%. The average SI value of SSIC is 56.34% which is very close to the average SI value of CMSA compensator. Though CMSA compensator shows better average SI results compared to SSIC, it is more complex. CMSA requires two Neutral codebooks, one for MFCC feature and the other for sinusoidal amplitude feature (SAF), for each speaker. In case of SSIC and other compensators, one MFCC feature based Neutral codebook is used. Hence, for average SI result point of view, SSIC can be a better choice over CMSA. The improvement in speaker recognition with a particular compensator is different for different emotions. It is observed that maximum SI result of 30.15% is obtained for Angry speech with speaker and stress information based compensator (SSIC) and combination of MFCC and SAF (CMSA) based compensators. For Lombard emotion, compensation by removal of stressed vector shows maximum SI result of 52.38%. Maximum SI result of 85.71% is obtained for Neutral emotion with CMSA compensator. Maximum SI result of 60.31% is obtained for Question speech with SSIC and CMSA compensators. Hence, an alternative solution to improvement in speaker recognition under stressed condition can be use of a compensator depending on the stressed condition or emotion. This requires prior determination of the stressed condition or emotion present in the speech signal.

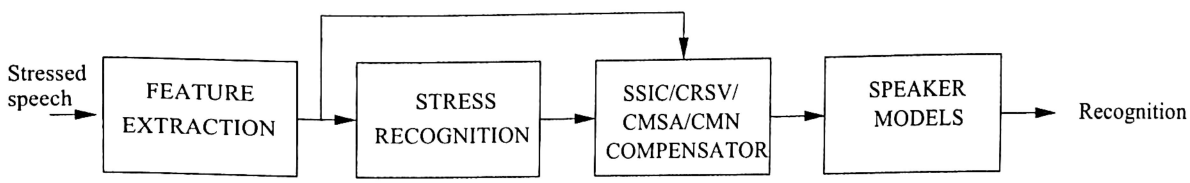


Figure 5.11: Block diagram of selection of compensation by Stress recognition (SCSR) system

Fig. 5.11 shows block diagram of selection of compensation by Stress recognition (SCSR) system. In this system, stress vectors are input to a stress recognition system. Then the compensator is chosen based on the recognition of stress condition. For Angry and Question emotion, SSIC is chosen as the compensator. CRSV is chosen for Lombard emotion and CMSA is chosen for Neutral emotion. In this work, the selection of a particular compensator is aimed at obtaining the best recognition results.

TH-1853_02610201

Table 5.8: SI results (%) of different compensators for SUSAS database

Compensators	Emotion of testing utterance				
	Angry	Lombard	Neutral	Question	Average
Un compensated	26.98	49.2	80.95	58.73	53.96
SSIC	30.15	50.79	84.12	60.31	56.34
CRSV	26.98	52.38	80.95	58.73	54.76
CMSA	30.15	50.79	85.71	60.31	56.74
CMN	26.98	41.5	84.84	58.73	52.95

Compensator selection will be different if other considerations such as complexity, real time application etc. are used. In the next subsection, we present stress recognition.

5.5.1 Stress recognition

Analysis and recognition of stressed speech is an active area of research since last two decades. Conventional speech features, which are widely used in various speech processing applications, are investigated to model stressed speech signals under different emotions. Nine speech features are used in this work for evaluation of their performance in stress or emotion recognition. These are mel-frequency cepstral coefficients (MFCC), three sinusoidal model based features and five linear prediction (LP) model based features. Sinusoidal model based features are sinusoidal frequency features (SFF), sinusoidal amplitude features (SAF) and sinusoidal phase features (SPF). The LP model based features are linear prediction cepstral coefficients (LPCC), log area ratios (LAR), arcsin reflection coefficients (ARC), reflection coefficients (RC) and linear prediction (LP) coefficients. Most of these features are widely used for speech and speaker recognition applications. LP coefficients are used to model stressed speech signal [94], [85]. In addition to LP coefficients, four LP derived features [2], cepstral coefficients estimated through LP coefficients (LPCC), log area ratios (LAR), reflection coefficients (RC) and arcsin reflection coefficients (ARC), are considered in this work for characterization of stressed speech. LP cepstral coefficients (LPCC) are used by various authors for characterization of stressed speech signal [105], [85], [98]. Though the other three features are not used for characterization of

TH-1853_02610201

stressed speech, they are extensively used for speech and speaker recognition applications. It would be useful to find their effectiveness in characterization of stressed speech. In this work, for all the above features, 12 coefficients are estimated from one speech frame as one feature vector.

Table 5.9: Stress classification using VQ classifier

Speech	Emotions				
Feature	Anger	Lombard	Neutral	Question	Average
MFCC	91.08	100	92.07	87.12	92.57
LAR	90.09	96.03	96.03	88.11	92.57
ARC	89.10	96.03	93.069	90.09	92.07
RC	91.08	92.07	89.10	90.09	90.59
LPCC	89.10	93.06	87.12	85.14	88.61
SAF	79.20	86.13	92.07	97.02	88.61
LP	80.19	82.17	74.25	71.28	76.98
SFF	86.13	79.20	66.33	72.27	75.99
SPF	78.21	74.25	75.24	76.23	75.99

The classification of stressed condition or emotion present in the signal is evaluated using the vector quantization (VQ) classifier. For training, we have considered 256 codewords for each emotion. The testing phase includes loading all the codebooks and then finding minimum distant codebook. A total of 1680 stressed speech data files are used for training and 840 stressed speech data files are used for testing. The results are shown in Table 5.9. Each column shows the recognition result of a particular emotion. It is observed that the performance of a particular feature varies for different emotions. Both MFCC and LAR features show maximum average recognition of 92.57%. LAR feature shows maximum recognition results with Neutral and Lombard emotions. ARC feature shows maximum recognition results with Lombard and Question emotion. RC feature shows maximum recognition results with Anger and Question emotion. Last column shows the average recognition performance of the features and they are shown in descending order. SPF feature shows the worst average recognition results.

Either MFCC or LAR feture can be chosen for stress recognition. For speaker recognition, it is shown that MFCC is the best feature. Hence, we choose MFCC feature for both stress and speaker recognition in a SCSR system shown in Fig. 5.11. For speaker identification using SCSR, CRSV compensator is chosen in the second stage if Lombard stressed condition is detected at the first stage.

TH-1853_02610201

Table 5.10: SI results (%) of different compensators for SUSAS database

Compensators	**Emotion of testing utterance**				
	Angry	Lombard	Neutral	Question	Average
Un compensated	26.98	49.2	80.95	58.73	53.96
SSIC	30.15	50.79	84.12	60.31	56.34
CRSV	26.98	52.38	80.95	58.73	54.76
CMSA	30.15	50.79	85.71	60.31	56.74
CMN	26.98	41.5	84.84	58.73	52.95
SCSR	30.15	52.38	85.71	60.31	57.14

If Neutral condition is detected CMSA technique is used. Otherwise, we select SSIC for speaker recognition under stressed condition. The speaker identification results for SCSR technique with the results obtained using the earlier techniques is shown in Table 5.10. The average speaker identification result has improved and its value is 57.14%. Specific properties of a particular stressed condition can be used for further improvement in speaker recognition in SCSR technique. In this work, we have used 16 size codebooks to demonstrate the degradations and improvements in speaker recognition under stressed condition in a visibly prominent manner.

TH-1853_02610201

Chapter 6

Conclusion

In this thesis, we investigate feature analysis and design of compensators for speaker recognition under stressed conditions. Stressed speech is induced by emotion, high workload, sleep deprivation, frustration and environmental noise. In stressed condition, the characteristics of speech signal are different from that of normal or neutral condition. Due to changes in speech signal characteristics, the performance of speaker recognition system degrades under stressed speech conditions. In this work, the problem of speaker recognition under stressed condition has been dealt with three broad approaches. First, six speech features (mel-frequency cepstral coefficients (MFCC), linear prediction (LP) coefficients, linear prediction cepstral coefficients (LPCC), reflection coefficients (RC), arc-sin reflection coefficients (ARC) and log-area ratios (LAR)), which are widely used for speaker recognition, are analyzed for evaluation of their characteristics under stressed condition. Statistical techniques such as Probability density function (pdf) characteristics, F-ratio test and Kolmogorov-Smirnov (K-S) test are used for evaluation of speaker discrimination capability of the features under stressed condition. Two classifiers, Vector Quantization (VQ) classifier and Gaussian Mixture model (GMM), are used to evaluate the speaker recognition results with different speech features. Relative ranking of the features based on their performance in speaker recognition under stressed condition is evaluated. This analysis help select the best feature set for speaker recognition under stressed condition. Secondly, sinusoidal model is used for analysis of speaker recognition system under stressed condition. Sinusoidal model has three different parameters or features. These are amplitude, frequency and phase features. With the assumption that these three features may carry different information of a speaker, the performance of the system can be improved by suitably combining the three features at the classifier stage. Finally, four novel compensation techniques are proposed and evaluated for improvement of speaker recognition under stressed condition. The compensation techniques are speaker and stressed information based

TH-1853_02610201

compensation (SSIC), compensation by removal of stressed vectors (CRSV), cepstral mean normalization (CMN) and combination of MFCC and sinusoidal amplitude (CMSA) features. Speech data from SUSAS database corresponding to four different emotions or stressed conditions, Angry, Lombard, Question and Neutral, are used for analysis of speaker recognition under stressed condition.

Normally, during the training sessions, speech signals are recorded in Neutral emotion condition. If the speech signal used for recognition is recorded in a different stressed condition, its characteristics may be different from that of the Neutral emotion condition. In Section 2.4, a 256-size VQ codebook is used for each speaker for evaluation of speaker identification under stressed condition. Four sets of codebooks, Angry, Lombard, Neutral and Question, corresponding to four emotions are implemented for each speaker. The testing utterances from each emotion are tested with the four sets of codebooks. It is observed that the SI results vary for the same codebook with test data from different emotions. Highest speaker recognition results for each emotion codebook is obtained when they are tested with the speech data of same emotion. These results prove that the speaker recognition performance is different if the training data and the test data are from different stressed conditions. The speaker recognition performance degrades if the test data are not from the same stressed condition. These results suggest that a detailed investigation is required for evaluation of performance of speaker recognition systems under stressed condition. During the last two decades, though some literature in the area of recognition of stressed speech are published, there is little effort for analysis and recognition of speakers under stressed conditions. Studies on speech signals produced under different stressed and noisy conditions have helped improve the speech recognition algorithms [84, 99, 100].

Six different speech features, LP coefficients, LPCC, ARC, RC, LAR and MFCC, are used for evaluation of speaker recognition under stressed condition. Different statistical feature evaluation techniques, such as Probability density characteristics, F-ratio test and Kolmogorov-Smirnov (K-S)test are used to assess the performances of the speech features. Vector quantization and Gaussian mixture model are used to assess the recognition performance. One of the goals of the analysis is to select features which are robust to stressed conditions. Relative ranking of the performance of different speech features will serve this purpose. Statistical analysis tools such as F-ratio test and K-S test provide insight into relative effectiveness of the coefficients of a feature in discriminating the speakers. Higher F-ratio value among the F-ratio values evaluated between the speakers indicates that the corresponding feature coefficient would work well for speaker recognition. Similarly, higher F-ratio value among the F-ratio values evaluated between the stressed conditions indicates that the feature coefficient would work well for stress recognition. It is observed that the F-ratio values between the speakers are different

TH-1853_02610201

from the F-ratio values between the stressed conditions. F-ratio value of eighth coefficient of MFCC feature has a relatively higher value compared to other coefficients when evaluated between the stressed conditions. Same eighth coefficient of MFCC feature has a relatively lower value compared to other coefficients when evaluated between the speakers. This result suggest that this coefficient is a good carrier of information related to stressed conditions and it is a poor carrier of speaker information. By suppression of this coefficient, the speaker recognition performance may improve because it will help eliminate the effect due to stress condition in a speaker's speech data. So, different coefficients can be weighted according to the applications for improvement in performance. This result proves that useful information for improvement in speaker recognition can be obtained by analysis of stressed conditions.

Relative ranking of the performance of different speech feature will serve this purpose. As speaker recognition is of primary interest in this work, the ranking of the features is carried out according to their recognition performances. Four types of speaker recognition experiments, VQ based speaker identification (SI), VQ based speaker verification (SV), GMM based SI and GMM based SV, are performed. MFCC, RC and ARC features have better performance compared to the other features. Similarly, LAR, LPCC and LP features show poor recognition results compared to other features in all the four recognition experiments. It is observed that the results of four speaker recognition experiments are different from each other. Under these circumstances, it is difficult to rank the features for their effectiveness in speaker recognition under stressed condition. We propose a method of ranking the features using cumulative probability values of relative performance of the features for each of the above experiment. MFCC feature is ranked as the best feature followed by RC, ARC, LAR, LPCC and LP features in descending order of their performances in speaker recognition under stressed condition.

Sinusoidal model based speech coder is well established for 4 kHz bandwidth transmission lines [118]. Sinusoidal model produces three sets of features. They are amplitude, frequency and phase features. Amplitude and frequency feature information are valuable for speaker recognition. These two features contain spectral peak information. The speaker information in these two features may be same or they may contain independent information. Speaker recognition can be improved by exploiting any independent or different speaker information present in these two features. In this work, we propose evaluation of confusion probability from confusion matrix for analysis of independent information with sinusoidal amplitude feature (SAF) and sinusoidal frequency feature (SFF). Results show that there may be considerable amount of independent information between the sinusoidal amplitude feature (SAF) and sinusoidal frequency feature (SFF). Improvement in speaker verification is observed by combining the amplitude and the frequency features.

TH-1853_02610201

The performance of the sinusoidal amplitude feature (SAF) is better than the frequency feature (SFF). But, the amplitude feature performance is relatively poor compared to MFCC. MFCC features are estimated using Fourier transform of mel-frequency band energies in the magnitude spectrum. The peak information in the spectrum is not directly used in case of MFCC feature as the spectral energy over a frequency band is used. Spectral peaks have speaker information. A combination of MFCC and sinusoidal features (SAF and SFF) can provide better speaker recognition performance. Speaker verification results with 106 male speakers' speech utterances from YOHO database are used for evaluation of verification results with the combination of features. It is observed that for combination of MFCC and SAF, the equal error rate is reduced from 0.55% to 0.36%. The reduction of percentage equal error rate, which shows improvement in speaker verification, proves that the sinusoidal amplitude feature (SAF) contains some speaker information which are not captured by MFCC feature. Comparison between the combination of SAF and MFCC and the combination of SFF and MFCC show slightly better result for combination of SAF and MFCC. Above idea of combining MFCC with SAF is extended to speaker recognition under stressed condition. It is observed that minimum equal error rate value is less than the equal error rate values of individual features for stressed speech.

It is observed that MFCC feature is the best feature for speaker recognition under stressed condition. In this work, the proposed compensation techniques are tested with MFCC features. Four novel compensation techniques are proposed and evaluated for improvement of speaker recognition under stressed condition. The compensation techniques are speaker and stressed information based compensation (SSIC), compensation by removal of stressed vectors (CRSV), combination of MFCC and sinusoidal amplitude (CMSA) feature based compensation and cepstral mean normalization (CMN).

Speaker and stress information based compensation(SSIC) technique is based on analysis of speaker and stress information using F-ratio values. It is observed that the F-ratio values between the speakers are different from the F-ratio values between the stressed conditions. Enhancing the effect of coefficients which have higher F-ratio values when evaluated between the speakers or reducing the effect of coefficients which have higher F-ratio values when evaluated between the emotions will result in improvement of performance of speaker recognition system. SSIC compensator is implemented at the decision stage of the codebook by evaluating a modified error distance. The modified error distance is estimated by dividing the codebook errors with the F-ratio values. This results in higher multiplication factor for the lower F-ratio valued coefficients. In this way the error distances for the lower valued coefficients are enhanced. It is observed that the speaker identification results for Angry and Question emotions are increased. The average speaker identification rate has improved to 56.34% from 52.77%

TH-1853_02610201

in case of uncompensated case.

Lower F-ratio valued coefficients have higher multiplication factors. Most of the lower valued F-ratio coefficients are higher order MFCC coefficients. Higher order MFCC coefficients are associated with source information [80]. Source is relatively more responsible for the emotion present in the speech signal compared to the vocal tract [129]. In the above experiment, only speaker discrimination information is used for finding the multiplication factors for the coefficients. Stress information is present with the F-ratio values evaluated between the stress conditions or emotions. Speaker recognition under stressed conditions can be improved if the stress information present with the F-ratio values of stress are properly exploited or used. It is observed that the average speaker identification rate has improved compared to the uncompensated case. For m2, its value is 57.13% and for m3, its value is 58.32%. With m2 multiplication factor the SI results are improved for Angry and Lombard emotions. Further improvements in Angry, Lombard and Question emotions are observed with the m3 multiplication factor. The improvements in SI results prove that best results can be obtained by suitably using both the speaker and stress information present in the signal.

Compensation by removal of stressed vectors (CRSV) is based on the idea of removing feature vectors which are most affected by the stressed condition. Stress effect on the features can be estimated by evaluating the distance between the feature vectors of the stressed speech and the Neutral speech. Feature vectors effected by stress are expected to be at larger distances compared to the feature vectors which are not effected by the stress condition. In this compensation technique, stressed feature vectors are detected by comparing the distance with a predefined threshold. Combination of MFCC and SAF (CMSA) features is proposed for improvement of speaker recognition under stressed condition. cepstral mean normalization (CMN) technique is traditionally used to compensate the noise for speech recognition [111]. In this work, CMN technique is tested for possible improvement in speaker recognition under stressed condition.

It is observed that different SI results are obtained for different stressed conditions or emotions. CMSA compensator shows maximum average SI value of 56.74%. The average SI value of SSIC is 56.34% which is very close to the average SI value of CMSA compensator. CMSA requires two Neutral codebooks, one for MFCC feature and the other for sinusoidal amplitude feature (SAF), for each speaker. In case of SSIC and other compensators, one MFCC feature based Neutral codebook is used. Hence, for average SI result point of view, SSIC can be a better choice over CMSA. The improvement in speaker recognition with a particular compensator is different for different emotions. It is observed that maximum SI result of 30.15% is obtained for Angry speech with speaker and stress information

TH-1853_02610201

based compensator (SSIC) and combination of MFCC and SAF (CMSA) based compensator. For Lombard emotion, compensation by removal of stressed vector shows maximum SI result of 52.38%. Maximum SI result of 85.71% is obtained for Neutral emotion with CMSA compensator. Maximum SI result of 60.31% is obtained for Question speech with SSIC and CMSA based compensators. Hence, an alternative solution to improvement in speaker recognition under stressed condition can be use of a compensator depending on the stressed condition or emotion. This requires prior determination of the stressed condition or emotion present in the speech signal.

The recognition of stressed condition or emotion present in the signal is evaluated using the vector quantization (VQ) classifier. For training, we have considered 256 codewords for each emotion. The testing phase includes loading all the codebooks and then finding minimum distant codebook. A total of 1680 stressed speech data files are used for training and 840 stressed speech data files are used for testing. It is observed that the performance of a particular feature varies for different emotions. Both MFCC and LAR features show maximum average recognition of 92.57%. LAR feature shows maximum recognition results with Neutral and Lombard emotions. ARC feature shows maximum recognition results with Lombard and Question emotion. RC feature shows maximum recognition results with Anger and Question emotion. SPF feature shows the worst average recognition results.

Either MFCC or LAR feture can be chosen for stress recognition. For speaker recognition, it is shown that MFCC is the best feature. Hence, we choose MFCC feature for both stress and speaker recognition in a SCSR system. For speaker identification using SCSR, CRSV compensator is chosen in the second stage if Lombard stressed condition is detected at the first stage. If Neutral condition is detected CMSA technique is used. Otherwise, we select SSIC for speaker recognition under stressed condition. The average speaker identification result has improved and its value is 57.14%. Specific properties of a particular stressed condition can be used for further improvement in speaker recognition in SCSR technique. In this work, we have used 16 size codebooks to demonstrate the degradations and improvements in speaker recognition under stressed condition in a visibly prominent manner.

6.1 Scope of future work

Some suggestions for future work with regard to the the problems studied by us are as follows.

1. Four stressed conditions are addressed in this work. Proposed analysis and compensation techniques can be extended and tested for more number of stressed conditions or emotions.

TH-1853_02610201

2. In this work, F-ratio values are used for evaluation of speaker information and stress information. More analysis is required for separation of speaker specific information from emotion or stress information present in the speech signal.

3. Perceptual models can be tested for evaluation of stress and speaker information.

4. Stress or emotion characteristics present in a signal depends on the language. Investigations can be taken up for analysis and design of compensators to take care of the above fact.

5. This work has addressed the problem of speaker recognition under stressed condition from the feature space point of view. Investigations can be taken up at the classifier level or speaker model point of view.

TH-1853_02610201

References

[1] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*. New Jersey: Prentice-Hall., 1978.

[2] L. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*. New Jersey: Prentice Hall Inc., Englewood Cliffs, 07632, 1993.

[3] G. Doddington, M. Przybockin, A. Martin, and D. Reynolds, "The NIST speaker recognition evaluation - overview, methodology, systems, results, perspective," *Speech Communication*, vol. 31, pp. 225–254, June 2000.

[4] B. S. Atal, "Automatic recognition of speakers from their voices," *Proc. IEEE*, vol. 64, pp. 460 – 476, April 1976.

[5] D. O. Shaughnessy, "Speaker recognition," *IEEE ASSP Mag.*, vol. 3, no. 4, pp. 4–17, Oct. 1986.

[6] J. Campbell, "Speaker recognition: A tutorial," *Proceedings of the IEEE*, vol. 85, pp. 1437 – 1462, Sept. 1997.

[7] F. Bimbot, J. F. Bonastre, C. Fredouille, G. G. I. Magrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-Garcia, D. Petrovska-Delacretaz, and D. Reynolds, "A tutorial on text-independent speaker verification," *EURASIP JASP*, vol. 4, pp. 430–451, 2004.

[8] S. Behera and S. Dandapat, "Speech features for speaker identification," *Journal of Institution of Engineers (Electronics and Telecommunications section) India*, vol. 83, pp. 1–6, July 2002.

[9] Furui, "Recent advances in speaker recognition," *Pattern Recognit. Lett.*, vol. 18, pp. 859 – 872, Apr. 1997.

[10] P. Sivakumaran, A. Ariyaeeinia, and M.J.Loomes, "Sub-band based text-dependent speaker verification," *Speech Communication*, vol. 41, pp. 485–509, October 2003.

TH-1853_02610201

[11] J. Markel, B. Oshika, and A. Gray, "Long-term feature averaging for speaker recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 25, pp. 330 – 337, Aug. 1977.

[12] S. D. J. Markel, "Text-independent speaker recognition from a large linguistically unconstrained time-spaced data base," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 27, pp. 74 – 82, Feb. 1979.

[13] A. E. Rosenberg and F. K. Soong, "Evaluation of a vector quantization talker recognition system in text independent and text dependent modes," *Computer Speech and Language*, vol. 2, pp. 143–157, September-December 1987.

[14] A. Poritz, "Linear predictive hidden markov models and the speech signal," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 7, pp. 1291 – 1294, May 1982.

[15] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital Signal Processing Academic Press*, vol. 10, pp. 19 – 41, January 2000.

[16] D. Reynolds, "Speaker identification and verification using gaussian mixture speaker models," *Speech Communication*, vol. 17, pp. 91 – 108, Aug. 1995.

[17] S. Furui, "Cepstral analysis technique for automatic speaker veri cation," *IEEE Trans. Acoust. Speech and Signal Process.*, vol. 29, p. 254272, 1981.

[18] S. Guruprasad, N. Dhananjaya, and B. Yegnanarayana, "AANN models for speaker recognition based on difference cepstrals," *Proceedings of the International Joint Conference on Neural Networks*, vol. 1, pp. 692 – 697, 20-24 July 2003.

[19] H. Misra, S. Ikbal, and B.Yegnanarayana, "Speaker-specific mapping for text-independent speaker recognition," *Speech Communication*, vol. 24, pp. 193–209, February 2003.

[20] M. Nosratighods, E. Ambikairajah, and J. Epps, "Speaker verification using a novel set of dynamic features," *18th International Conference on Pattern Recognition*, vol. 4, pp. 266 – 269, 20-24 Aug. 2006.

[21] R. D. Zilca, B. Kingsbury, J. Navratil, and G. N. Ramaswamy, "Pseudo pitch synchronous analysis of speech with applications to speaker recognition," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, pp. 467–478, March 2006.

TH-1853_02610201

[22] C. S. Liu, W. J. Wang, M. T. Lin, and H. C. Wang, "Study of line spectrum pair frequencies for speaker recognition," *International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 277 – 280, 3-6 April 1990.

[23] B. Imperl, Z. Kacic, and B. Horvat, "A study of harmonic features for the speaker recognition," *Speech Communication*, vol. 22, pp. 385–402, Sept. 1997.

[24] S. Umesh, L. Cohen, and D. Nelson, "Frequency warping and the mel scale," *IEEE Signal Processing Letters*, vol. 9, pp. 104 – 107, March 2002.

[25] S. Umesh, L. Cohen, and D. Nelson, "Frequency-warping and speaker-normalization," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, pp. 983 – 986, 21-24 April 1997.

[26] S. Umesh, L. Cohen, and D. Nelson, "Fitting the mel scale," *IEEE International Conference on Acoustics, Speech, and Signal Processing,*, vol. 1, pp. 217 – 220, 15-19 March 1999.

[27] C. H. Liu and O. T. C. Chen, "A text-independent speaker identification system using parcor and ar model," *The 2002 45th Midwest Symposium on Circuits and Systems*, vol. 3, pp. III–332 – III–335, 4-7 Aug. 2002.

[28] K. S. R. Murty and B. Yegnanarayana, "Combining evidence from residual phase and MFCC features for speaker recognition," *IEEE Signal Processing Letters*, vol. 13, pp. 52–55, Jan. 2006.

[29] Shi-Han Chen and Hsiao-Chuan Wang, "Improvement of speaker recognition by combining residual and prosodic features with acoustic features," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. I–93–96, 17-21 May 2004.

[30] S. Kim, T. Eriksson, Hong-Goo Kang, and Dae Hee Youn, "A pitch synchronous feature extraction method for speaker recognition," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. I – 405–408, 17-21 May 2004.

[31] S. Kajarekar, L. Ferrer, A. Venkataraman, K. Sonmez, E. Shriberg, A. Stolcke, H. Bratt, and R. Gadde, "Speaker recognition using prosodic and lexical features," *IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 19 – 24, 30 Nov.-3 Dec. 2003.

TH-1853_02610201

[32] J. S. Mason and X. Zhang, "Velocity and acceleration features in speaker recognition," *1991 International Conference on Acoustics, Speech, and Signal Processing*, vol. 5, pp. 3673 – 3676, 14-17 April 1991.

[33] A. Abdelatty Ali, J. Van der Spiegel, and P. Mueller, "An acoustic-phonetic feature-based system for the automatic recognition of fricative consonants," *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, pp. 961 – 964, 12-15 May 1998.

[34] B. Yegnanarayana, K. S. Reddy, and S. Kishore, "Source and system features for speaker recognition using AANN models," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 409 – 412, 7-11 May 2001.

[35] P. Sivakumaran and A. M. Ariyaeeinia, "The use of sub-band cepstrum in speaker verification," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, pp. II1073 – II1076, 5-9 June 2000.

[36] A. Kinney and J. Stevens, "Wavelet packet cepstral analysis for speaker recognition," *Conference Record of the Thirty-Sixth Asilomar Conference on Signals, Systems and Computers*, vol. 1, pp. 206 – 209, 3-6 Nov. 2002.

[37] C. T. Hsieh, E. Lai, and Y. C. Wang, "Robust speech features based on wavelet transform with application to speaker identification," *IEE Proceedings-Vision, Image and Signal Processing*, vol. 149, pp. 108 – 114, April 2002.

[38] Ka-Yee Leung, Man-Wai Mak, and Sun-Yuan Kung, "Applying articulatory features to telephone-based speaker verification," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. I – 85–88, 17-21 May 2004.

[39] Sung-Wan Yoon, S. Kim, and Hong-Goo Kang, "A supplementary feature for speaker verification in constrained room acoustics," *Proceedings of 2004 International Symposium on Intelligent Signal Processing and Communication Systems*, pp. 189 – 192, 18-19 Nov. 2004.

[40] F. Farahani, P. Georgiou, and S. Narayanan, "Speaker identification using supra-segmental pitch pattern dynamics," *IEEE International Conference on Acoustics, Speech, and Signal Processing, 2004. Proceedings.*, vol. 1, pp. I – 89–92, 17-21 May 2004.

TH-1853_02610201

[41] N. M. H. Hermansky, "Speaker verification using speaker specific mapping," *in Proc. of Speaker Recognition and its Commercial and Forensic Applications, France*, pp. 111–14, April 1998.

[42] B. Yegnanarayana, S.R.M. Prasanna, J.M. Zachariah, and C.S. Gupta, "Combining evidence from source, suprasegmental and spectral features for a fixed-text speaker verification system," *IEEE Transactions on Speech and Audio Processing*, vol. 13, pp. 575 – 582, Jul. 2005.

[43] I. M. Chagnolleau and G. Durou, "Application of time-frequency principal component analysis to text-independent speaker identification," *IEEE Trans. on Speech and Audio Processing*, vol. 10, pp. 371–378, Sept. 2002.

[44] Ravi P. Ramachandran, R. Kevin, Farrell, Roopashri Ramachandran, and Richard J. Mammone, "Speaker recognition-general classifier approaches and data fusion methods," *Pattern Recognition*, vol. 35, pp. 2801–2821, December 2002.

[45] T. Quatieri, *Discrete-Time Speech signal processing.* Pearson education (Singapore) Pte.Ltd, 2004.

[46] Y. Zhang, C. Loke, R. Togneri, and M. Alder, "A comparison of PBDHMM and CHMM for isolated word recognition," *Proc. Fifth Australian International Conference on Speech Science and Technology, Perth, Australia*, pp. 564–569, December 1994.

[47] M.Inal, E.Butun, K.Erkan, M.Yildirum, and C.Ceken, "Comparison of linear predictive analysis methods for ANN-based speaker identification," *Proc. IEEE 5th seminar on neural network applications in electrical engineering*, pp. 109–112, Sept. 2000.

[48] L. Bryan, Pellom, and J. Hansen, "An efficient scoring algorithm for gaussian mixture model based speaker identification," *IEEE Signal Processing Letters*, vol. 5, pp. 281–284, Nov. 1998.

[49] K. R. Farrell, R. J. Mammone, and K. T. Assaleh, "Speaker recognition using neural networks and conventional classifiers," *IEEE Transactions on Speech and Audio Processing*, vol. 2, pp. 194–205, Jan. 1994.

[50] Y. Linde, A. Buzo, and R. M. Gray, "An algorithm for vector quantizer design," *IEEE Transactions on Communications*, vol. 28, pp. 84–95, Jan. 1980.

[51] T. Kohonen, "The self-organizing map," *IEEE Proc.*, vol. 78, p. 1464 1480, Sept. 1990.

TH-1853_02610201

[52] J. He, L. Liu, and G. Palm, "A discriminative training algorithm for VQ-based speaker identification," *IEEE Transactions on Speech and Audio Processing*, vol. 34, p. 353 356, May 1999.

[53] I. Booth, M. Barlow, and B. Watson, "Enhancements to DTW and VQ decision algorithms for speaker recognition," *Speech Communication*, vol. 13, pp. 427–433, December 1993.

[54] P. C. Chang and B. H. Juang, "Discriminative template training for dynamic programming speech recognition," *IEEE International Conference on Acoustic, Speech and Signal Processing, San Francisco, California*, pp. I–493I–496, March 1992.

[55] J. Wilpon and L. Rabiner, "A modified K-means clustering algorithm for use in isolated word recognition," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 33, pp. 587–594, Jun. 1985.

[56] S. C. H. Sakoe, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 26, p. 43 49, Feb. 1978.

[57] F. Itakura, "Minimum prediction residual principle applied to speech recognition," *IEEE Trans. Acoust. Speech and Signal Process.*, vol. 23, p. 67 72, Feb. 1975.

[58] S. Furui, "Comparison of speaker recognition methods using statistical features and dynamic features," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 29, pp. 342–350, Jun. 1981.

[59] T. K. Moon, "The expectation-maximization algorithm," *IEEE Signal Proc. Mag*, vol. 13, p. 47 60, Nov. 1996.

[60] D. Reynolds, "Comparison of background normalization methods for text-independent speaker verification," *European Conference on Speech Processing, Rhodes, Greece*, p. 963 966, September 1997.

[61] D.A. Reynolds and R.C. Rose, "Robust text-independent speaker identification using gaussian mixture speaker models," *IEEE Trans. Speech Audio Process.*, vol. 3, pp. 72 – 83, Jan. 1995.

[62] D. Reynolds, "The effects of handset variability on speaker recognition performance: Experiments on the switchboard corpus," *IEEE International Conference on Acoustic, Speech and Signal Processing, Atlanta, Georgia*, pp. 113–116, May 1996.

TH-1853_02610201

[63] R.C. Rose, E.M. Hofstetter, and D.A. Reynolds, "Integrated models of signal and background with application to speaker identi cation in noise," *IEEE Trans. Speech Audio Process.*, vol. 2, pp. 245 –257, April 1994.

[64] G. P. M. Saon, G.; Zweig, "Linear feature space projections for speaker adaptation," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 325 – 328, 7-11 May 2001.

[65] J. Picone, "Continuous speech recognition using hidden markov models," *IEEE Acoust. Speech Signal Process. Mag.*, vol. 7, pp. 26–41, July 1990.

[66] T. Matsui and S. Furui, "Comparison of text-independent speaker recognition methods using VQ distortion and discrete/contuniuous hmms," *IEEE International Conference on Acoustic, Speech and Signal Processing, San Francisco, California*, pp. II–157–II–160, March 1992.

[67] L. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, pp. 257–286, Feb. 1989.

[68] T. Matsui and S. Furui, "Concatenated phoneme models for text-variable speaker recognition," *IEEE International Conference on Acoustic, Speech and Signal Processing, Minneapolis, Minnesota*, pp. II–391–II–394, April 1993.

[69] O. Siohan, C. H. Lee, A. C. Surendran, and Q. Li, "Background model design for fexible and portable speaker verification systems," *IEEE International Conference on Acoustic, Speech and Signal Processing*, pp. 825–828, March 1999.

[70] J. Oglesby and J. Mason, "Optimisation of neural models for speaker identification," *International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 261 – 264, 3-6 April 1990.

[71] K. R. Farrell, "Text-dependent speaker verification using data fusion," *International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 349 – 352, 9-12 May 1995.

[72] X. Yue, D. Ye, C. Zheng, and X. Wu, "Neural networks for improved text-independent speaker identification," *IEEE Engineering in Medicine and Biology Magazine*, vol. 21, pp. 53 – 58, Mar.-Apr. 2002.

TH-1853_02610201

[73] M. W. Mak and S. Y. Kung, "Estimation of elliptical basis function parameters by the EM algorithm with application to speaker verification," *IEEE Transactions on Neural Networks*, vol. 11, pp. 961 – 969, July 2000.

[74] W. Ren-hua, H. Lin-shen, and H. Fujisaki, "A weighted distance measure based on the fine structure of feature space: Appliction to speaker recognition," *Proc. ICASSP*, pp. 273–276, 1990.

[75] H. Gish and M. Schmidt, "Text-independent speaker identification," *IEEE Signal Processing Magazine*, vol. 11, pp. 18 – 32, Oct. 1994.

[76] A. E. Rosenberg and S. Parthasarathy, "Speaker background models for connected digit password speaker verification," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 81–84, 7-10 May 1996.

[77] H.Jiang and L.Deng, "A Bayesian approach to the verification problem: Applications to speaker verification," *IEEE Trans. Speech Audio Process*, vol. 9, pp. 883–884, Nov. 2001.

[78] K. P. Markov and S. Nakagawa, "Text-independent speaker recognition using non-linear frame likelihood transformation," *Speech Communication*, vol. 39, pp. 301–310, June 1998.

[79] R. Aukenthaler, M. Carey, and H. Thomas, "Score normalization for text-independent speaker verification systems," *Digital siganl processing*, vol. 10, pp. 42–54, Jan. 2000.

[80] D. O. Shaughnessy, *Speech Communication: Human and Machine*. Addison-Wesley, Reaing, MA, 1987.

[81] G. S. Raja and S. Dandapat, "Sinusoidal model based speaker identification," *Proc. NCC-2004 conference, IIsc, Bangalore*, vol. 1, pp. 523–527, Jan.-Feb. 2004.

[82] G. S. Raja and S. Dandapat, "Sinusoidal model based speaker identification using VQ and DHMM," *Proc. IEEE, INDICON-2004 conference, IIT, Kharagpur*, pp. 338–343, Jan.-Feb. 2004.

[83] A. E. Rosenberg, "Automatic speaker verification: A review," *IEEE Proceedings of the IEEE*, vol. 64, no. 4, pp. 475–487, 1976.

[84] J. H. L. Hansen and D. A. Cairns, "ICARUS: Source generator based real-time recognition of speech in noisy stressful and lombard effect environments," *Speech communication*, vol. 16, pp. 391 – 422, June 1995.

TH-1853_02610201

[85] S.Ramamohan and S.Dandapat, "Sinusoidal model-based analysis and classification of stressed speech," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, pp. 737 – 746, May 2006.

[86] Hayes, "Statistical digital signal processing and modeling," *New York: John Wiley and sons, Inc.*, vol. 17, 1996.

[87] L. Bosch, "Emotions, speech and the ASR framework," *Speech Communication*, vol. 40, pp. 213–225, April 2003.

[88] C. M. Lee and S. S. Narayanan, "Toward detecting emotions in spoken dialogs," *IEEE Trans. Speech and Audio Processing*, vol. 13, pp. 293–303, March 2005.

[89] R. Cowie, E. D. Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor, "Emotion recognition in human-computer interaction," *IEEE Signal Processing Magazine*, vol. 18, pp. 32–80, Jan. 2001.

[90] B. Schuller, G. Rigoll, and M. Lang, "Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine - belief network architecture," *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 577–580, 2004.

[91] J. Tao, Y. Kang, and A. Li, "Prosody conversion from neutral speech to emotional speech," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, pp. 1145–1154, July 2006.

[92] K. R. Scherer, "Vocal communication of emotion: A review of research paradigms," *Speech Communication*, vol. 40, pp. 227–256, April 2003.

[93] D. Ververidis and C. Kotropoulos, "Emotional speech recognition: Resources, features and methods," *in Press Speech Communication*, 2006.

[94] S. Ghazale and J. H. L. Hansen, "A comparative study of traditional and newly proposed features for recognition of speech under stress," *IEEE Trans. Speech and Audio Processing*, vol. 8, pp. 429–442, July 2000.

[95] B. Zhou and J. H. L. Hansen, "Rapid discriminative acoustic model based on eigenspace mapping for fast speaker adaptation," *IEEE Trans. Speech Audio Process.,*, vol. 13, pp. 554–564, Jul. 2005.

TH-1853_02610201

[96] T. L. Nwe, S. W. Foo, and L. C. D. Silva, "Classification of stress in speech using linear and nonlinear features," *Proc. IEEE ICASSP*, pp. II.9–II.12, Apr. 2003.

[97] H. Patro, G. S. Raja, and S. Dandapat, "Classification of stressed speech using gaussian mixture model," *Proc. IEEE Indicon*, pp. 342–346, Dec. 2005.

[98] R. Sarikaya and J. H. L. Hansen, "High resolution speech feature parametrization for monophone-based stressed speech recognition," *IEEE Signal Processing Letters*, vol. 7, pp. 182–185, July 2000.

[99] Y. Gong, "Speech recognition in noisy environments: A survey," *Speech Commun.*, vol. 19, pp. 261–291, April 1995.

[100] J. H. L. Hansen and M. A. Clements, "Source generator equalization and enhancement of spectral properties for robust speech recognition in noise and stress," *IEEE Trans. Speech Audio Process.*, vol. 3, pp. 407–415, Sep. 1995.

[101] S. E. Bou-Ghazale and J. H. L. Hansen, "A comparative study of traditional and newly proposed features for recognition of speech under stress," *IEEE Trans. Speech Audio Process.*, vol. 8, pp. 429–442, Jul. 2000.

[102] J. Picone, "Signal modeling techniques in speech recognition," *Proc. IEEE*, vol. 81, pp. 1215–1247, Sep. 1993.

[103] J. H. L. Hansen and B. L. Pellom, "Text-directed speech enhancement employing phone class parsing and feature map constrained vector quantization," *Speech Commun.*, vol. 21, pp. 169–190, April 1997.

[104] I. Shahin, "Enhancing speaker identification performance under the shouted talking condition using second-order circular hidden markov models," *Speech Communication*, vol. 48, pp. 1047–1055, August 2006.

[105] J. H. L. Hansen and B. D. Womack, "Feature analysis and neural network-based classification of speech under stress," *IEEE Trans. Speech and Audio Processing*, vol. 4, pp. 307–313, July 1996.

[106] M.Hunt, "Further experiments in text-independent speaker recognition over communications channels," *Proc .IEEE Intern. Conf. ASSP*, pp. 563–566, Apr. 1983.

TH-1853_02610201

[107] S. Mathur and B. H. Story, "Vocal tract modeling: implementation of continuous length variations in a half-sample delay Kelly-Lochbaum model," *Proceedings of the 3rd IEEE International Symposium on Signal Processing and Information Technology, Germany*, pp. 753 – 756, Dec. 2003.

[108] D. A. Cairns and J. H. L. Hansen, "Nonlinear analysis and classification of speech under stressed conditions," *The Journal of the Acoustical Society of America*, vol. 96, pp. 3392–3400, December 1994.

[109] J. H. L. Hansen, B. Womack, and L. M. Arsian, "A source generator based production model for environmental robustness in speech recognition," *International Conference on Spoken Language Processing (ICSLP)*, pp. 1003– 1006, 1994.

[110] S. E. Bou-Ghazale and J. H. L. Hansen, "Stress perturbation of neutral speech for synthesis based on hidden markov models," *IEEE Trans. Speech and Audio Processing*, vol. 6, pp. 201–216, May 1998.

[111] D. Roe, "Speech recognition with a noise-adapting codebook," *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '87*, vol. 12, pp. 1139 – 1142, Apr. 1987.

[112] Y. Chen, "Cepstral domain stress compensation for robust speech recogniton," *IEEE intrenational coference on Acoustics, Speech, and Signal Processing, ICASSP '87*, vol. 12, no. 717-720, 1987.

[113] J. Makhoul, "Linear prediction: A tutorial review," *Proceedings of the IEEE*, vol. 63, pp. 561 – 580, April 1975.

[114] R. Tucker, A.J. Robinson, J. Christie, and C. Seymour, "Recognition-compatible speech compression for stored speech," *Proc. Accessing information in spoken audio*, pp. 69–72, April 1999.

[115] R. Viswanathan and J. Makhoul, "Quantization properties of transmission parameters in linear predictive systems," *IEEE Trans on acoustics, speech and signal processing*, vol. ASSP-23, pp. 309–321, June 1995.

[116] W. Press, S. Teukolsky, W. Vetterling, and Flannery, "Numerical recipes in C," *Cambridge: Cambridge University*, 1992.

TH-1853_02610201

[117] F. Bimbot, J. F. Bonastre, C. Fredouille, G. Gravier, M. I. Chagnolleau, S. Meignier, T. Merlin, O. J. Garcia, P. Delacretaz, and Reynolds, "A tutorial on text-independent speaker verification," *EURASIP J. Applied Signal process.*, vol. 2004, no. 4, pp. 430-451, 2004. doi:10.1155/S1110865704310024.

[118] H.Jiang and L.Deng, "A Bayesian approach to the verification problem: Applications to speaker verification," *IEEE Trans. Speech Audio Process*, vol. 9, pp. 883–884, Nov. 2001.

[119] Y. Stylianou, "A simple and fast way of generating a harmonic signal," *IEEE signal processing letters*, vol. 7, pp. 111–113, May 2000.

[120] E. George and M. Smith, "Speech analysis/synthesis and modification using an analysis-by-synthesis/overlap-add sinusoidal model," *Speech and Audio Processing, IEEE Transactions*, vol. 5, pp. 389 – 406, Sept. 1997.

[121] J. Jensen and J. H. L. Hansen, "Speech enhancement using a constrained iterative sinusoidal model," *IEEE Trans. Speech and Audio Processing*, vol. 9, pp. 731–740, October 2001.

[122] C. Etemoglu, V. Cuperman, and A. Gersho, "Speech coding with an analysis-by-synthesis sinusoidal model," in *ICASSP*, vol. 3, pp. 1371–1374, June 2000.

[123] T. Quatieri and R. Danisewicz, "An approach to co-channel talker interference suppression using a sinusoidal model for speech," *IEEE Transaction on Acoustics, Speech, and Signal Processing*, vol. 38, pp. 56–69, Jan. 1990.

[124] L.S.Marques and L. Almeida, "Frequency-varying sinusoidal modeling of speech," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, pp. 763–765, May 1989.

[125] R. McAulay and T. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Trans. Acoustics, Speech, and Signal processing*, vol. 34, pp. 744 – 754, Aug. 1986.

[126] B. Friedlander and J. Francos, "Estimation of amplitude and phase parameters of multicomponent signals," *IEEE Transactions on Signal Processing*, vol. 43, pp. 917–926, April 1995.

[127] G. S.Raja and S. Dandapat, "Stress compensation for improvement in speaker recognition," *Proceedings of IEEE INDICON, India Habitat Centre, Lodhi Road, New Delhi-110003, India*, pp. 1–5, Sep. 15-17 2006.

TH-1853_02610201

[128] J. H. L. Hansen, "Analysis and compensation of speech under stress and noise for environmental robustness in speech recognition," *Speech communication*, vol. 20, pp. 151–173, November 1996.

[129] K. Rao and B. Yegnanarayana, "Prosody modification using instants of significant excitation," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, pp. 972 – 980, May 2006.

[130] D.G. Childers and C. F. Wong, "Measuring and modeling vocal source-tract interaction," *IEEE Transactions on Biomedical Engineering*, vol. 41, pp. 663–671, July 1994.

TH-1853_02610201

Related Publications by the Author

Refereed Journals

1. G. S. Raja, and S. Dandapat, "Performance of Selective Speech Feature for Speaker Identification", *Accepted (Nov-2006) in Institution of Engineers (India) Journal.*

2. Ramamohan. R., G. S. Raja, H. B. Nemade and S. Dandapat., "Online Speech Recognition for Medical Transcription", *Revised and submitted to IEEE Transactions on Biomedical Engineering.*

3. H. Patro, G. S. Raja, and S. Dandapat, "Statistical Feature Evaluation for Classification of Stressed Speech", *Revised and submitted (August-2006) to International Journal of Speech Technology.*

4. G. S. Raja, and S. Dandapat, "Comparison of speech features for the recognition of stressed speech", *Communicated to IEEE Transactions on Audio, Speech and Language Processing.*

Manuscripts Under Preparation:

1. G. S. Raja, and S. Dandapat, "Compensation for Speaker Recognition under Stressed Condition", *to be submitted to IEEE Transactions on Audio, Speech and Language Processing.*

2. G. S. Raja, and S. Dandapat, "Combination of Sinusoidal Model and MFCC Information for Speaker Recognition", *to be submitted to Speech Communication.*

Conference Publications:

1. G. S. Raja, and Dandapat, S., "Sinusoidal Model and MFCC Information for Speaker Recognition", *Accepted for IEEE TENCON 2007*, October 30- November 2, 2007, Taipei, Taiwan.

2. G. S. Raja, and S. Dandapat, "Stress compensation for improvement in speaker recognition", in *Proceedings of IEEE INDICON 2006*, India Habitat Centre, Lodhi Road, New Delhi-110003, India, pp. 1 - 5, Sep. 15-17, 2006.

3. H. Patro, G. S. Raja, and S. Dandapat, "Classification of Stressed Speech using Gaussian Mixture Model" in *Proceedings of Annual IEEE international conference named IEEE INDICON 2005*, Indian Institute of Technology Madras, India, pp. 342-346, Dec. 11-13, 2005.

TH-1853_02610201

4. G. S. Raja, and S. Dandapat, "Stressed speech analysis for ergonomic design" in *Proceedings of International Ergonomics Conference*, Indian Institute of Technology Guwahati, India, pp.251-255,Dec. 10-12, 2005.

5. G. S. Raja and S. Dandapat, "Sinusoidal model based speaker identification using VQ and DHMM" in *Proceedings of IEEE First India annual conference named IEEE INDICON 2004,* , Indian Institute of Technology Kharagpur, India, pp.338 - 343, Dec. 20-22, 2004.

6. G. S. Raja, and S. Dandapat, "Sinusoidal Model Based Speaker Identification", in *Proceedings of NCC-2004*, IISc, Bangalore, India, pp.527-531, Jan. 30 - Feb. 1, 2004.

TH-1853_02610201

CURRICULUM VITAE

1. NAME:

   G.SENTHIL RAJA

2. DATE OF BIRTH:

   18 January 1977

3. EDUCATIONAL QUALIFICATIONS:

   - April-1999 B.E
   - December-2000 M.E
   - July-2007 Thesis submitted for Ph.D degree

4. PERMANENT ADDRESS:

s/o G.Gunaseela Boopathy
868c, Thukkapet (Post), Chengam (Taluk)
Thiruvannamalai (District)
Tamil nadu (State), India
606709
Ph: +91-04188-222282

TH-1853_02610201

Response to Examiner's Queries

Indian Examiner

Pg. 19, 2nd para, 3rd line: Spelling error in "mixture".

We have corrected the word "Mmixture" to "Mixture".

Pg. 22, 2nd para, 5th line: 3.5-72-41: Should it be 35-72-41?

" 3.5-72-41" is corrected as "35-72-41"

Pg. 22, 2nd para, last line: Standard telephone bandwidth is 3.1 kHz (300-3400 Hz) and not 3.8 kHz.

3.8 kHz is replaced by 3.1 kHz (300-3400 Hz).

Pg. 35, 4th line, "Speaker verification.... is somewhat more difficult than speaker identification"; In general, in speaker verification problem, the system is presented with a claimed identity that needs to be accepted or rejected, whereas in speaker identification, a speaker has to be identified from a set of speakers without the benefit of a claimed identity. I would consider a speaker identification task to be more difficult rather than the other way round. Any comments?

Yes. Speaker identification is more complicated than speaker verification. It has been corrected in place.

Pg. 37, tables 2.1, 2.2: What is the dimension of each of the feature vectors?

12 coefficients are used for YOHO and SUSAS database. This is mentioned in the revised thesis.

TH-1853_02610201

Pg. 39, 2nd line.: "Coefficients one, two ... have their histogram peaks shifted from third bin to fourth bin between the emotions Neutral and Angry". Should it be "shifted from fourth bin to third bin" as suggested by the figure?

Yes. We have modified the histogram in the revised thesis.

Pg. 40, 1st para, 5th last line: "Highest recognition result is obtained for Angry and Lombard trained data with Angry and Lombard test data respectively." However, Neutral trained data with Neutral test data shows higher recognition than the angry case in Table 2.4.

There was mistake in the experimentation process for obtaining results tabulated in Table 2.4. The mistake was the feature vectors of angry emotion were appended to Lombard emotion feature vectors while training the Lombard model. The same thing was happening in training neutral emotion model as well as question emotion model. Hence, the results tabulated were wrong. The experiment is redone now and the correct results are tabulated. In view of this, the observations about the results are also modified accordingly in the thesis.

Pg. 40, Table 2.4: The author may explain why Neutral training –Neutral testing case is poorer that Lombard training – Lombard testing case, as this is not intuitively expected.

This experiment has been checked thoroughly and found an issue like the previous emotion training vectors are appended with the following training vectors. The results are corrected in thesis.

Pg. 44, 2nd para, 2nd line: "The recognition results in Section 2.3 are based on the Neutral speech codebook."Can you also confirm that the results of Table 2.2 (SUSAS) are also based on the neutral codebook?

Yes. Table 2.2 results are obtained by considering only neutral speech data from SUSAS database. This is mentioned in the revised thesis.

TH-1853_02610201

Pg. 50, last sentence of 2nd para: "Features extracted from pre-emphasized speech contain better speaker discrimination information compared to features obtained from non pre-emphasized speech which is evident from the average F-ratio shown in Table 3.1." This is not always true for neutral speech from the table. This statement should be appropriately qualified.

Yes. In case of LP and LPCC features, the above statement is not correct for neutral speech case. Except these two, in all other cases the pre-emphasized features are showing higher F-ratio values. Accordingly, the above statement is modified in the revised thesis as follows:

"In majority of the cases (excluding LP and LPCC for neutral speech), features extracted from pre-emphasized speech contain better speaker discrimination information compared to features obtained from non pre-emphasized speech which is evident from the average F-ratio shown in Table 3.1".

Pg. 56: Table 3.4: Have the linear prediction coefficients been used as such, without any transformation? Usually distances between LP coefficients correlate poorly with spectral envelope distances, and are not used.

We agree with this point. To observe the recognition performance under stressed condition we have conducted an experiment directly using LP coefficients. However, for this case also it results very poor performance.

Pg 76-77: Speaker discrimination with sinusoidal features: (a) Amplitude feature will depend on the energy level of the speaker. Have all the speakers' energy been normalized in some manner (such as over the sentence) before using them for speaker discrimination? (b) Do the speakers speak the same sentences, otherwise the energy contours which are dependent on the sentence transcription, will be different, and should not be compared. (c) Has any dynamic time warping been done to scale the duration of utterances?

Energy normalization has been used to maintain the same energy level. Yes. The speakers speak the same sentences. To analyse the speaker recognition under stressed condition with natural time duration, dynamic time warping is not used. However, this aspect is taken care by the compensation techniques.

Pg. 80: Figure 4.3: Why is the phase varying only between -0.3 and 0.35?

It's a warped phase. Speaker recognition performance has been tested with warped phase only. This part of the work was motivated from the good performance by warped phase for stress classification. However, significant benefit may be achieved by using unwarped phase.

Pg. 83: It will be useful to provide a quantitative comparison of sinusoidal feature based speaker recognition with other classical features based methods.

The results for other classical features are given in Table 2.1. In the revised thesis, this table is reproduced as Table 4.x and compared with proposed sinusoidal features.

Pg. 84, 85 : Rows and columns of C_SAF, C_SFF,P_SAF, P_SFF need to be defined.

In C_SAF, C_SFF, P_SAF and P_SFF, the columns represent the speaker models and the rows represent the speaker of test utterances. This information is given in the revised thesis.

Pg. 87, table 4.2: When you mention the stressed condition, does it exclude the neutral condition in the averaging?

No. In the present work, neutral emotion is also considered as stressed condition. Hence, all the averaged results include neutral condition. However, even excluding the neutral condition provides same relative improvement in the performance.

TH-1853_02610201

Pg. 97, 4th line: "This results in higher multiplication factor for the lower F-ratio valued coefficients." One approach is to give more weight to those features which are likely to provide more robust recognition performance i.e. give more weight to larger F-ratios. (This would be similar to the Maximal Ratio Combiner for equalizers in communications). Comments on the rationale for your approach.

This approach is a F-ratio weighted distance computation. In distance computation, more emphasis is given to the coefficients having high F-ratio values. This provides improvement in the performance as demonstrated.

Pg 101: 5th last line: "During testing, each test vector is compared with the CNC vector. If the distance between these two vectors is more than the threshold the tested vector is retained." If feature vectors greater than a threshold distance are retained, then are the stressed vectors being retained or removed?

Assumption here is the stressed vectors would be more distanced than non stressed vectors from common neutral code book and has to be removed. But it's observed that the performance improvement in either way stressed vectors are retained. As well as, the performance improvement is observed only for Lombard. So, we have concluded that the speaker information is more in speech part which is having Lombard effect than the other part of the speech.

TH-1853_02610201

Foreign Examiner

The candidate used SUSAS and YOHO database for analysis. The consistency of the usage needs to be addressed. The candidate has frequently shifted from use of one database to another without justifying the shift. SUSAS represents a speech under stress corpus (which was collected by me for research on speech recognition under stress back in the 1980's, and released to the LDC in 2000). YOHO is a well respected speaker recognition corpus. SUSAS has a single recording session - and therefore does not represent session-to-session variability, a CORE issue all researchers in speaker recognition question if a single session is only used. Also, for tests on the stress conditions from SUSAS, there are only 70 words (2 tokens for each of 35 words). If some of these are used for training, this leaves a limited number of test tokens for evaluation. To address the consistency issue in the document: the candidate can overcome this issue by focusing on results from a particular database (in this case, YOHO as it does not contain stress component) and then expand his results for SUSAS database. However, it will be a challenge to obtain acceptance of the results with SUSAS for speaker ID since there is just too small a number of speakers with too limited speaker materials.

All our methods are initially tested using YOHO database. This is due to the availability of comparable results in the literature from Reynold's and Cambell's works. Whenever analysis needs to be done with reference to stressed condition, then results are obtained for SUSAS database. We completely agree that SUSAS database is a very small one for speaker recognition task. Since, there was no other standard stressed speech database, we used this one. We have also collected our own database with 48 speakers, since SUSAS has less number of speakers. In our database, four emotions namely, Neutral, Angry, Question and Lombard have been considered. For each emotion, we recorded 20 utterances. 15 utterances are used for training and 5 utterances are used for testing. The experimental results for the same are enclosed as appendix in the revised thesis. A copy of the same is enclosed here for your kind reference.

TH-1853_02610201

The number of coefficients for a particular feature (either MFCC, or other features) is not consistent (or sometimes not clear). On some occasions the candidate uses 12 coefficients, on other occasions, 14 coefficients, or some other values without a suitable justification. Many researchers argue that if when an increased size feature set is used, greater discrimination ability should be achieved and a direct comparison between features is not as fair. A strongly suggestion would be that the candidate focus on use of "same" number of coefficients for a feature set across all the experiments (which would require the candidate to re-run the experiments again). If this is not possible, at least including a justification of the choice of the feature dimension size and nothing these when discussing the results would help temper the objection from a reviewer on an external publication.

Throughout the work, we have considered 12 coefficients. Only in compensation analysis 14 coefficients are considered. This is with the objective of removing about two coefficients which are affected significantly from stress. This information is clearly mentioned in the revised thesis at appropriate places. However, for reference and uniformity, results of fourteen coefficients case are shown for 12 coefficients given in table 5.4.

The candidate needs to evaluate his algorithm for a corpus which has sentences (SUSAS has isolated words; under stress it is expected that a consistent level of stress/emotion is not conveyed across a sentence versus a single word); alternatively the candidate should include his comments on algorithm performance for such a scenario? Why were these four stress conditions selected? Especially the "question" style. If the candidate has attempted his algorithm for other stress categories in SUSAS, what were the results for other stress categories? Can the candidate comment on what would be the results for other stress categories if the candidate has not performed experiments?

Now, a 48 speaker database has been used to evaluate all the methods in this thesis. The results are included at the end of the thesis as appendix. We picked randomly four emotions including neutral from SUSAS database for speaker recognition under stressed condition. There was no specific reason in choosing question style. We have not conducted any experiments for other stressed conditions. However, the loud stress condition is similar to that

TH-1853_02610201

of Lombard stress. Hence, performance for this stress condition may be similar to that of Lombard stress. On the similar lines the soft, slow and clear stress conditions may have performance nearly same as that of neutral stress condition.

In general, can the candidate comments on whether the results from the various experiments are significant or not? For example, in Table 3.5/pg. 58: was ARC coefficients feature set significantly better as compared to the second best feature set? The speaker ID rates for Angry for example range between 22-33%, with three features achieving 31-33%. How would the reader know which of the three are statistically better than the other? The number of test tokens from SUSAS is simply too small for a confidence measure to be obtained here.

Since the number of speakers as well as testing data is small, it is very difficult to comment on statistical significance of the results. The reliability of the results may be verified by testing the proposed methods on a larger database. In this work now all the methods have been tested on relatively larger database of 48 speakers as mentioned above. The results on this database also show a similar trend. From this we can say that the observations made on the proposed methods using SUSAS database are reliable.

Page 6) If Joe Campbell has provided a comparison of different features such as RC, LAR, ARC, LSP frequencies, and MFCC, what is gained by repeating the same experiments as part of the thesis? The idea is to check the performance across STRESS conditions - so it is necessary to ensure your performance on neutral is consistent with what was reported in Campbell. Was this comparison made? (I did not find it here?).

We are extremely sorry for this statement made in the thesis. Campbell in his paper has provided only a description about different features that can be used for speaker recognition. Based on this observation we explored all of them for stressed condition. We have modified the corresponding statement in the revised thesis as follows:

"Campbell [6] has provided a description of different features such as reflection coefficients, log area ratios, arc-sin reflection coefficients, line spectrum pair (LSP) frequencies and mel-frequency cepstral coefficients (MFCC)."

Page 8) why was only the unsupervised methods (VQ, GMM) considered in this thesis? Since all the speech data in SUSAS is known words (of the +12000 tokens, about 100 are documented errors/repeats – these were tokens that were not collected of obtained properly and the SUSAS read me file explains there). Supervised speaker recognition would have been a reasonable area to explore – perhaps with background noise if that was of interest. Can you comment on not using supervised methods for this thesis?

There was no specific reason for not exploring supervised methods. Otherwise, the usage of VQ and GMM is based on the availability of literature and resources.

Page 8) what specific type of VQ was implemented in the thesis? There are several, not limited to LBG method, LVQ methods, GVQ method? How many classes were used for clustering in the VQ? The selection of the VQ codebook size is a sensitive parameter for performance in speaker identification.

Binary split algorithm has been used in our work. Different sizes of codebook are tested and shown the results for 16 and 256 in thesis. Most of the results are shown for codebook size 16 to compare with GMM. GMM is restricted up to 16 mixtures because of insufficient data.

Page 9) in the GMM, did the candidate adapt (either MAP or MLLR or combination) the speaker models? It is not clear if a sufficient amount of data exists to construct speaker models while performing testing. If a "round-robin" solution was employed- train with 90% of the data, test with 10%. Repeat until all data has been used for testing – a maximum amount of data can be leveraged to get results. This means much more computation. Please elaborate if this was done (if so, I missed in the document).

During training no adaptation is done. A common VQ codebook is created and its centroid vectors are used as initial weight vectors for training GMMs. Since there is no similar

database for creating UBM, no adaptation technique is employed. 60% (42 out of 70) of the data is used for training and 10% (7 out of 70) is used for testing. This is mentioned in the thesis (pp. 40).

Page 28) Looking at the PDF plots for RC, LAR, ARC, I do not see much of a difference across these, as these features are based on LP coefficients. Why did the candidate choose these? Can the candidate have all the plots in same scale? (It is very bad to plot these on different scales since the reader cannot compare differences easily). Also, with only 9 speakers, why were only 4 used here?

The plots have all been redrawn to the same scale. For better visualization, four out of nine are randomly chosen for plotting.

Page 30) same conclusion as above from F-test values.

In the revised thesis the plots have all been redrawn to the same scale.

Page 31) abbreviation CDF used before explaining it.

Abbreviation is properly defined before its usage in the revised thesis.

Page 32) the candidate has tested the KS-test on YOHO database which does not have stressed speech. How can he extend these results to test scores to the SUSAS database? Or is the focus here to just report the results and not extend the results to the SUSAS corpus?

YOHO database is used to illustrate how the KS-test can be used. Based on this illustration, the KS-test parameters are recomputed afresh for SUSAS database.

Page 34) For the text after Equation 2.3.2: Please refer the following PowerPoint presentation:... On slide 14, the text used in the thesis is has been used verbatim from this

page. I personally do not know the original source, but clearly it is not the candidate's words. I include copy here.

The proper reference is cited. In the revised thesis, it is [117].

Page 35) inconsistent writing ... the writing is disjoint.

We have corrected the sentences in that particular page to maintain continuity.

Page 41) "It is observed that the features which carry speaker dependent information are "effected" due to the stress condition. The word "effected" is not correct/appropriate." Should be impacted or affected.

We have corrected the word "effected" to "affected".

Page 41) last sentence of the page (continued on to next page) why does the text refer to AUTHORS (plural)?

We have corrected the particular sentence in the revised thesis.

Page 47-48) Do figures 3.2 and 3.3 tell a different story versus what is claimed by the candidate in the thesis? It appears so.

The motivation for this part of the work is that if the mean values of the coefficients are different in the pdfs, then they may have discriminatory information. This may be with reference to either speaker or emotion. This aspect is pictorially illustrated in the Figures 3.2 and 3.3. From these figures it can be observed that the mean values across different speakers are different in some coefficients (c1, c4). Thus these coefficients have better discrimination across the speakers. Similar discussion can be extended for other coefficients and also emotions as given in the thesis.

Page 63) Please specify the "units" used for the results here. So that when the value "higher" is used for better SI, and lower the better for SV will be clear. It appears represent accuracy (identification accuracy), so units in % accuracy, and SV is EER.

In the revised thesis, throughout the thesis the units are mentioned as correct identification rate in % for speaker identification and EER in % for speaker verification.

Though throughout the thesis report, the candidate works with pre-emphasis and non-pre-emphasised signal; it would be helpful to mention it again here that the analysis looks dispersed (and not at all concise/consistent between sections/parts of the study).

We agree that there is some inconsistency in using various concepts at different places. This is mainly because all these concepts are independently explored and while including them in thesis, the unified view was difficult to achieve. We are sorry about the same.

The idea is appealing; probably the candidate wants to highlight the good points about his feature set.

Yes. These experiments are aimed to show the sinusoidal model feature performance better than other features.

For some experiments (Page 57, last paragraph) the candidate chooses 16 sizes VQ, and then for SI and SV, he makes a choice of 256 size code book (page 59) why? This needs justification.

GMM needs sufficient data to train the speaker models but VQ does not require that much data. Using SUSAS database we could increase the number of mixtures up to 16 for GMM. To compare the GMM based results we used same size of code book for VQ. Finally, VQ based 256 size codebook has been used to show the better performance than 16 size codebook.

TH-1853_02610201

Page 64) the candidate selects 16 mixtures in his GMM modelling, why? Did he try some other combination? Implying that was it from experimentation or some theory-heuristic based decision? For speaker ID, researchers normally use 32 mixtures when an energy threshold is used to remove low energy frames. What was the energy threshold for silence/low energy threshold is used, then 16 is picked because there is just too little data to train and therefore the evaluations should be supervised to be meaningful.

Due to limited data we directly considered only 16 mixtures.

Page 66) Ranking the features based on their performance seems to be an odd goal here; however because the candidate is addressing the credibility of the feature-sets across all the four tasks, then it makes sense. However, can there be some different techniques to rank order these in a meaningful way - some statistical approach to mitigate this issue? The technique used seems to be appropriate" for the task.

There may be better ways of rank ordering these features. Currently we are unable to explore anyone.

Page 73) Frame dropping concept – frames at the beginning and end are removed if they contain energy less than `10% of the total energy. This is fair for SUSAS, since about 1/3 to 50% of the waveform samples represent leading and training silence for the isolated words. This "10%" is a 0.1 normalized threshold- so there should be more mixtures in the GMM if there is sufficient amount of speech (43 phonemes represented by 16 mixtures? Not likely to be sufficient to discriminate the speakers).

Since, we have used only 60% of the data for training GMM, we have chosen only 16 mixtures.

Page 74) Why the square of peak values? (Log is understandable – to reduce the dynamic range of the value). What is the FFT size here? The FFT size is mentioned on page 80 (last line), but it would nice to have it mentioned here as well. The consistency in the use of a feature set – with respect to the choice of feature and the feature size is missing. Sometime

TH-1853_02610201

(page 87, 74, and elsewhere), the candidate uses 12- coefficients, elsewhere log square and sometimes, the feature itself.) The candidate should stick to use of a consistent feature set. As well, the candidate never tries to tie up the results from YOHO database to SUSAS database, because of this size inconsistency. The suggestion would be compile results on two different database separately and then compare the results for different feature sets with respect to the same data base [and then across the two databases, if required].

It was a mistake. We have considered log of the peak values, not the square of the peak values. This is corrected in the revised thesis. To bring some unified view what is done in the revised thesis is to add a new appendix. In this appendix, all the results related to each database are tabulated separately.

Page 78] as the candidate wants to compare the Sinusoidal features with the traditional features, can he please regenerate the plots in figure 4.2, 4.3 for 12 th order sinusoidal model? Otherwise section 4.1 and 4.2 seems to be discontinued.

The plots have been regenerated and included in the respective sections.

Page 82] Explain: The identification rate is improved by pitch dependent speaker identification rate because of removal of spurious speaker" –in context of the content in this paragraph.

The statement is wrong and not required in the context. Alternatively, this statement is included in that place "The motivation for using this statement in that context is to convey that by using some minimum frequency difference between successive peaks, the spurious ones can be eliminated".

Page 83] Results described on YOHO database needs to be tabulated to give a comprehensive – high level view of the work. Furthermore, extra analysis on figure 4.12 will help. Can these results be extended to other database? Can you cite other studies that use YOHO and compare your performance with theirs?

TH-1853_02610201

The baseline speaker recognition results for this thesis is obtained using YOHO database. In the literature, earlier results exist for same database in Reynold's work [16]. Tables 2.1 and 2.3 show the overall speaker identification and verification results for YOHO database. Analyses with Reynold's results are discussed on page number 38. Same experiments are extended for sinusoidal model and reported here. Finally we extended it in to speaker recognition under stressed condition. The experimental results for different studies using YOHO database have been included in Appendix B.

[16] D. Reynolds, "Speaker identification and verification using gaussian mixture speaker models," Speech Communication, vol. 17, pp. 91 – 108, Aug. 1995.

Page 85) the candidate has represented the differences in the information content with different metrics.

We have represented the confusion matrix of SAF and SFF in terms of probability values. However, as mentioned these two matrices represent the same information.

Page 89) the remark-"Overall speaker identification rate with SFF feature is comparable with the speaker identification results of SAF feature" – is incorrect.

Agreed and removed this line from the thesis.

Page 90) though still, SAF and MFCC are closely related, except for the DCT of the filter bank energies. Please modify the text to reflect this.

The text has been modified in the revised thesis to reflect this. We have added the following statements to take care of this.

"However, it would be emphasized at this point that still SAF and MFCC are closely related due to their similar origin. The difference between these two is SAF has peak information and MFCC is obtained by using discrete cosine transform (DCT) on the filter bank energies."

TH-1853_02610201

Page 98] interpretation form Figure 5.3 seems unnecessary

Figure 5.3 and its reference in the text are removed from the revised thesis.

Page 98] the technique describing the compensation scheme is nice one. By having a lower multiplication factor [0.2] for the stress component, the performance is better as compared to that from higher multiplication factor [0.5]. This is an interesting conclusion and may require more analysis from the candidate. Again, the candidate used 14 coefficients to highlight his point. He then changes to 12 MFCC coefficients. The reason behind this difference needs to be explained.

Initially we have analysed with 14 coefficients. Final results are shown with 12 MFCC coefficients.

Table 5.4 and other tables and figures in chapter 5] please include the actual value of the multiplicative factor.

Yes. The multiplication factor values are included in figure and table.

Page 102] please rephrase the sentence "It is observed that the recognition for Lombard speech has improved resulting in overall improvement of average recognition which is 54.76%." Please specify the improvement in Lombard section only, (either absolute or relative %) and then specify the overall improvement.

From the table it can be observed that the performance improves only in the Lombard speech case from 49.2% to 52.38%. As a result, the overall improvement of average recognition performance increases from 53.96% to 54.96%.

Page 105] The result of using alpha =0.65 for unknown prior (on stress condition) is understandable (intuitive) because, expect for Lombard stress, all the other stress conditions

are more-or-less constant across the variation of alpha, and that Lombard stress performs the best for that value.

Page 103-107] this is a nice analysis, where the candidate 's contribution can be clearly seen and results match the analysis. Not much of analysis done here with the GMM based modelling technique. The focus is on VQ classifier, which again leads me to believe that the mixture size/training data was not sufficient to represent the speaker accurately. The VQ based method actually lends itself nicely to supervised based speaker ID, yet the focus here is on unsupervised (the extremely low ID performance form Table 5.6 could be improved if it were supervised! And give the candidate a better chance to impress reviewers on a journal/ conference paper).

Thank you very much for this suggestion. Due to some constraints we are unable to implement the supervised cases. However, this is included as scope for future work in the revised thesis.

****Typographic error*
Page 73] Closing parenthesis missing – (frame with the length of 20ms...
Page 85]f[R]ist- character Missing.
Page 86] evaluated-appears twice.

We have corrected all these typographic errors.

Page 103]"But, amplitude feature performance is comparatively poor with MFCC." Does the candidate want to say that SAF+MFCC performance to be poor as compared to MFCC alone, or MFCC features perform better as compared to SAF features?

It is meant to mean that MFCC features perform better as compared to SAF features. However SAF+MFCC performance is better when they are combined in a weighted manner with 0.35 for SAF and 0.65 for MFCC for SUSAS database.

Page 103]"Above idea of combining MFCC and SAF can is extended to speaker recognition under stress condition" – can and is?

Corrected as "MFCC and SAF can be …."

Section 5.3] needs rewriting.

We have rewritten the section 5.3 in the revised thesis.

Page 103] Table 5.6 seems to be hanging. It's explained on paper 105. It would be nice to add the value of alpha in the stress column for which the max performance was obtained.

It is corrected in revised thesis. Alpha value has been included in table.

TH-1853_02610201

Appendix - A

To evaluate our algorithms and compensation techniques, 48 Speaker Database (40 males and 8 females) Under Stressed Condition (SDUSC) has been used in this appendix. In this database, twenty utterances of four emotions from forty eight speakers are collected. From these utterances, fifteen utterances per speaker have been used for training and five utterances per speaker have been used for testing. This data base gives comparatively high speaker recognition than SUSAS data base. Because this data base is collected is in a single session.

Feature analysis for speaker recognition under stressed condition

In feature analysis, 12 coefficients are extracted and used for training and testing. Each row in Table 1 represents the code book emotions and each column the testing emotions respectively. Each value shows result for the speaker recognition value for testing utterances under emotion indicated by column and for the code book used indicated by rows. The observation shows that the recognition results are different for the same code book with test data from different emotions. Also maximum recognition results are obtained diagonally. From overall performance the average value is higher for neutral trained code book.

Table 1: Speaker identification rate for different emotions

Training (16 size code book)	Testing				
	Angry	Lombard	Neutral	Question	Average
Angry	93.75	83.33	82.50	80.83	85.10
Lombard	89.29	94.44	82.54	81.42	86.87
Neutral	87.50	84.58	95.00	90.83	89.48
Question	89.58	90.83	84.58	93.33	89.58

Table 2 shows the performance of different features when evaluated by four different stress recognition experiments. The recognition results of eight speech features are shown in descending order by their performances. MFCC, LPCC and LAR features gives better performance compared to other features. Similarly, RC, ARC and LP features show poor

TH-1853_02610201

recognition when compared to other features. It is observed that the results are different from each other under these circumstances it is difficult to rank the features for their effectiveness in speaker recognition under stressed condition. We suggest the following approach for ranking the features. The probabilities of their positions in the four columns are evaluated from this table. Each feature appears once in each column each position in a row is assigned a probability value of 0.25 for a feature if it appears. LAR, LPCC, ARC and MFCC features appears once in first row. So, they are assigned a probability value of 0.25 for first ranking. MFCC appears twice in the second row. So, it is assigned probability value of 0.5. Similarly, probability values of the features according to their positions are estimated and are shown in the Table 3. Each row in this table shows the ranking probabilities of the features. Final ranking of features is decided based on the cumulative probability. Cumulative values are shown in Table 4. Cumulative probability of MFCC attained the value of one by the third row. LPCC and LAR features attain the value of one by the fifth row. Hence these three features qualify for the top three rankings.

Table 2: Feature analysis for speaker recognition under stressed condition (Pre-emphasized)

Rank	VQ based Speaker identification rate (%)	VQ based Speaker Verification (EER in %)	GMM based speaker identification rate (%)	GMM based speaker verification (EER in %)
1	LAR (90.42)	MFCC(0.0604)	ARC(94.58)	LPCC(0.054)
2	MFCC (89.58)	LPCC(0.11)	RC(91.25)	MFCC (0.051)
3	ARC (88.75)	SAF(0.16)	MFCC(85)	ARC(0.081)
4	LPCC (84.17)	RC(1.01)	LAR(85)	RC(0.95)
5	SAF (83.34)	LAR(1.12)	LPCC(81.25)	LAR (0.973)
6	RC (71.67)	ARC(2.06)	LP(77.92)	SAF(1.56)
7	LP(64.58)	SFF(2.12)	SAF(69.58)	LP(1.803)
8	SFF(51.67)	LP(2.19)	SFF(62.92)	SFF(1.934)

TH-1853_02610201

Table 3: Position probability of six speech features for speaker recognition under stressed condition

Position	MFCC	LPCC	LAR	ARC	RC	SAF	LP	SFF
1	0.25	0.25	0.25	0.25	0	0	0	0
2	0.5	0.25	0	0	0.25	0	0	0
3	0.25	0	0	0.5	0	0.25	0	0
4	0	0.25	0.25	0	0.5	0	0	0
5	0	0.25	0.5	0	0	0.25	0	0
6	0	0	0	0.25	0.25	0.25	0.25	0
7	0	0	0	0	0	0.25	0.5	0.25
8	0	0	0	0	0	0	0.25	0.75

Table 4: Position probability of six speech features for speaker recognition under stressed condition

Position	MFCC	LPCC	LAR	ARC	RC	SAF	LP	SFF
1	0.25	0.25	0.25	0.25	0	0	0	0
2	0.75	0.5	0.25	0.25	0.25	0	0	0
3	1	0.5	0.25	0.75	0.25	0.25	0	0
4	0	0.75	0.5	0.75	0.75	0.25	0	0
5	0	1	1	0.75	0.75	0.5	0	0
6	0	0	0	1	1	0.75	0.25	0
7	0	0	0	0	0	1	0.75	0.25
8	0	0	0	0	0	0	1	1

TH-1853_02610201

Stress Compensation for speaker recognition

MFCC feature has been considered for the all of the following compensation techniques. Because the feature analysis shows that MFCC performs better than other features.

1. Combination of MFCC and sinusoidal amplitude features

Here we are going to compare the equal error rate of amplitude feature and Mel-frequency cepstral coefficients. Equal error rate of amplitude feature is 0.16 % and for Mel-frequency cepstral coefficient 0.0604%. Individually the weight for the amplitude feature is 0.04 and for MFCC is 0.96. Combining the weight of the two features, the equal error rate is 0.0543. Fig. 1 shows the equal error rate for different weights between amplitude feature and Mel-frequency cepstral coefficients.

Fig.2 shows Genuine (diamond symbol) and Imposter (circular symbol) score of combination of sinusoidal model amplitude feature and Mel-frequency cepstral coefficients. Five testing utterances per speaker are considered to evaluate the combination score (amplitude, Mel-frequency feature) for speaker recognition. So the number of genuine score is 240(48 speaker x 5 utterances). The number of imposed score is 11280(47 imposed score x 48 speaker x 5 utterances).

Table 5: Equal Error Rate of amplitude feature and Mel-frequency cepstral coefficient

Amplitude based EER in %	Mel-frequency cepstral coefficient based EER in %	Combination of Amplitude and Mel-frequency cepstral coefficient based EER
0.16	0.0604	0.0543 (amplitude feature weight is 0.04, MFCC feature weight is 0.96)

TH-1853_02610201

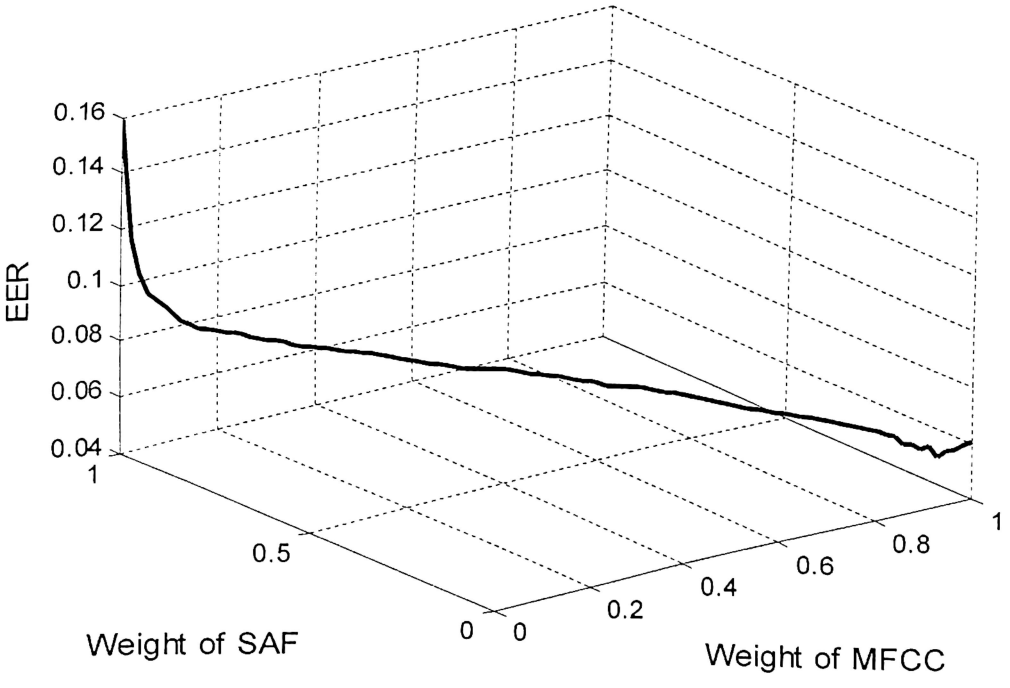


Fig. 1 Equal error rate of MFCC and SAF for stressed speech

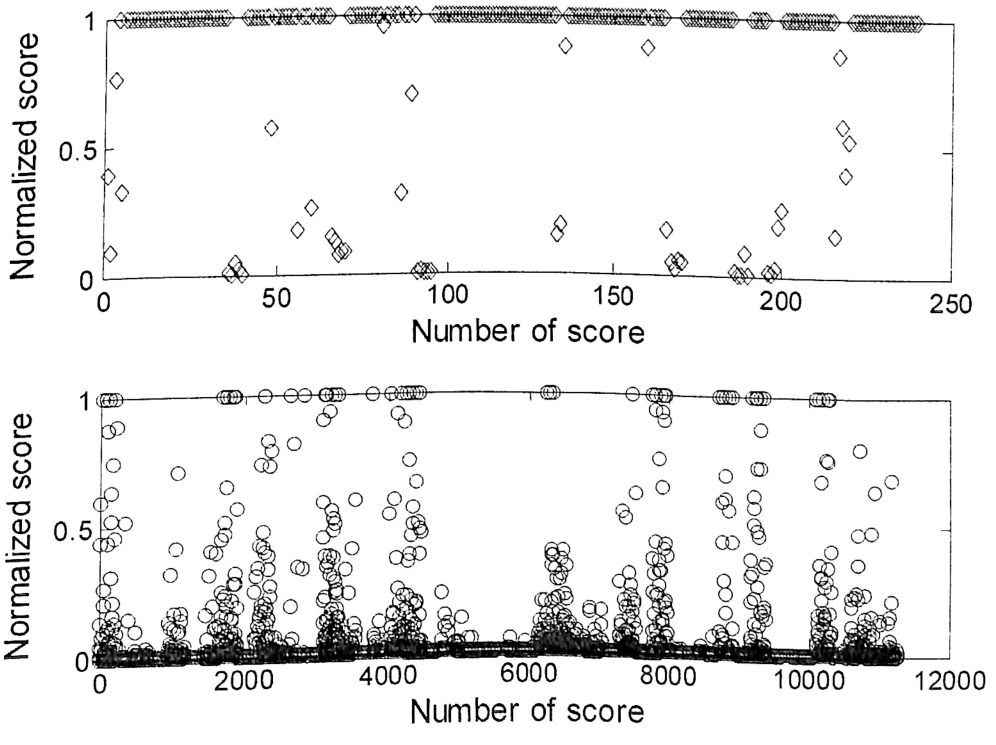


Fig. 2 Genuine (diamond symbol) and Imposter (circular symbol) score of combination of sinusoidal model amplitude feature and Mel-frequency cepstral coefficients

TH-1853_02610201

Table 6 : Correct Identification Rate of amplitude feature and Mel-frequency cepstral coefficient

Compensator	Emotion of testing utterance				
	Angry	Lombard	Neutral	Question	Average
Uncompensated	87.50	84.58	95.00	90.83	89.47
Combined MFCC and SAF (amplitude feature weight is 0.04, MFCC feature weight is 0.96)	91.25	85.83	97.50	92.91	91.87

2. Speaker and Stress Information based Compensation (SSIC)

The distance between codebook vector and testing utterance vector are weighted by inverse F-ratio value of each coefficient. In table 7, m1 represent the F-ratio values of speakers then finding the reciprocal of those values. Here, 0.25 times the mean of F-ratio values evaluated between the four stress conditions is added to the F-ratio values of the speakers and then the reciprocal of these values are taken as m2.

Table 7: Speaker identification rate (%) for different multiplication factors with the error distances in the code book.

Compensators/ Multiplication factors	Testing				
	Angry	Lombard	Neutral	Question	Average
Un compensated	87.50	84.58	95.00	90.83	89.47
m1	92.91	85.00	93.75	92.50	91.04
m2	93.53	89.42	93.17	91.52	91.97

TH-1853_02610201

3. Compensation by Removal of Stressed Vectors (CRSV)

It is observed that in the 48 speakers also the Lombard emotion gives improved performance which is 89.43 %

Table 8: Speaker identification rate (%)

Training (16 size code book)	Testing				
	Angry	Lombard	Neutral	Question	Average
Un compensated	87.50	84.58	95.00	90.83	89.47
Stressed vector removal	86.25	89.58	95.00	90.41	90.31

4. Cepstral Mean Normalization

Table 9: Speaker identification rate (%)

Training (16 size code book)	Testing				
	Angry	Lombard	Neutral	Question	Average
Un compensated	87.50	84.58	95.00	90.83	89.47
Cepstral mean normalisation (Neutral training utterances)	87.50	77.50	98.33	90.41	88.43
Cepstral mean normalisation (All emotion training utterances)	85.00	82.50	98.897	98.75	88.33

TH-1853_02610201

5. Selection of compensation by stress recognition (SCSR)

In this compensation technique, stress recognition has been used to select one of the above compensation techniques. MFCC feature has been used for both the speaker recognition and stress recognition. The speaker recognition degrades for neutral emotion testing because compensation applied for neutral condition also.

Table 10: SI results (%) of different compensators

COMPENSATORS	Testing				
	Angry	Lombard	Neutral	Question	Average
Un compensated	87.50	84.58	95.00	90.83	89.47
Speaker and Stress Information based Compensation	93.33	89.58	93.33	91.66	91.97
Stressed vector removal	86.43	89.58	95.00	90.41	90.31
Combined MFCC and SAF	91.25	85.83	97.5	92.91	91.87
Cepstral mean normalisation	85	82.5	98.75	87.08	88.33
Selection of compensation by stress recognition	93.75	90.41	94.16	93.16	92.91

From the experimental results of SUSAS and 48 speaker databases, it is confirmed that the recognition results are different for the neutral codebook with test data from different emotion. The proposed combined compensation technique improves the performance of speaker recognition system under all stressed conditions. This confirms that the proposed compensation techniques work in a similar manner irrespective of population (Number of speakers).

TH-1853_02610201

6. Conclusions

In this Appendix, forty eight speakers database under stressed condition has been used to analyse the performance of different speech features to observe the performance degradation. Eight different speech features like, LPCs, LPCC, ARC, MFCC, LAR, RC, SAF and SFF are used for evaluation of speaker recognition under stressed condition. The results show that MFCC feature results better recognition compare to others. Similar performance is obtained in case of SUSAS database. In the next step, different compensators are evaluated to observe the improvement in the recognition performance. The proposed combined compensation technique improves the performance of speaker recognition system under all stressed conditions as in the case of SUSAS database.

TH-1853_02610201

Appendix – B

In this appendix, the different speech features are compared with different databases. Speaker identification results of neutral speech are tabulated with three different databases in Table 1, Table 2 and Table 3. In speaker identification, comparatively MFCC feature shows better recognition than other features. From these three tables, the performance of sinusoidal amplitude features show better recognition than LP features except SUSAS database.

1. Speaker Identification Results under Neutral Conditions

Table 1: Speaker identification results of YOHO database (106 male speakers, 12 coefficients, 256 size codebook) for different speech features.

Speech features	MFCC	LAR	LPCC	ARC	RC	LP	SAF	SFF
Correct Identification rate (%)	98.49	97.64	96.42	89.62	87.74	86.42	79.91	44.06

Table 2: Speaker identification results of neutral speech in SUSAS database (9 speakers, 12 coefficients, 32 size codebook) for different speech features.

Speech features	MFCC	LAR	ARC	RC	LPCC	LP	SFF	SAF
Correct Identification rate (%)	84.13	80.95	79.37	79.37	77.77	73.02	65.07	57.14

TH-1853_02610201

Table 3: Speaker identification results of neutral speech in SDUSC database (48 speakers, 12 coefficients, 32 size codebook) for different speech features.

Speech features	MFCC	LPCC	LAR	ARC	RC	SAF	LP	SFF
Correct Identification rate (%)	95.00	93.75	90.83	89.16	87.08	74.16	70.41	52.08

2. Speaker Verification Results under Neutral Conditions

In speaker verification, the equal error rates of different speech features for YOHO, SUSAS and SDUSC database are shown in Table 4, Table 5 and Table 6, respectively. MFCC feature shows least equal error rate than all other listed speech features. SUSAS database shows higher equal error rate than other two databases because of limited database.

Table 4: Speaker verification results of YOHO database (106 male speakers, 10 background speakers, 12 coefficients and 256 size codebook) for different speech features

Speech features	MFCC	LPCC	LAR	ARC	RC	SAF	LP	SFF
Equal error rate (%)	0.55	0.57	0.71	0.83	0.83	1.25	2.92	37.92

Table 5: Speaker verification results of neutral speech in SUSAS database (9 speakers, 12 coefficients, 32 size codebook) for different speech features

Speech features	MFCC	LAR	ARC	RC	LPCC	LP	SAF	SFF
Equal error rate (%)	4.58	6.88	8.12	8.12	9.16	12.60	14.37	27.50

Table 6: Speaker verification results of neutral speech in SDUSC database (48 speakers, 12 coefficients, 32 size codebook) for different speech features

Speech features	MFCC	LPCC	LAR	ARC	RC	SAF	LP	SFF
Equal error rate (%)	0.04	0.06	0.09	0.092	0.094	0.12	0.30	10.72

3. Speaker Identification Results under Stressed Conditions

Speaker identification results under stressed condition are shown in Table 7 and Table 8. The speaker identification rate of MFCC is lesser than ARC in SUSAS database. If we compare Table 7 and Table 8, MFCC feature performs better than all other features in position rank wise.

Table 7: Speaker identification results of SUSAS database (9 speakers, 12 coefficients, 32 size codebook) for different speech features

Speech features	ARC	MFCC	RC	LAR	LPCC	LP	SFF	SAF
Speaker identification rate (%)	55.55	54.76	52.77	50.79	49.20	49.60	37.30	36.50

Table 8: Speaker identification results of SDUSC database (48 speakers, 12 coefficients, 32 size codebook) for different speech features

Speech features	MFCC	LPCC	LAR	ARC	RC	SAF	LP	SFF
Speaker identification rate (%)	89.58	84.17	90.42	88.75	71.67	83.34	64.58	51.67

TH-1853_02610201

4. Speaker Verification Results under Neutral Conditions

Speaker verification results are shown in Table 9 and Table 10 for stressed speech databases. Here, comparatively MFCC feature shows less equal error rate than other features. SDUSC database collected in a single session as well as more data so that it shows less equal error rate than SUSAS database.

Table 9: Speaker verification results of SUSAS database (9 speakers, 12 coefficients, 32 size codebook) for different speech features

Speech features	MFCC	SAF	RC	ARC	LAR	LPCC	SFF	LP
Equal error rate (%)	9.06	10.36	12.84	13.47	16.20	18.58	21.12	23.79

Table 10: Speaker verification results of SDUSC database (48 speakers, 12 coefficients, 32 size codebook) for different speech features

Speech features	MFCC	LPCC	SAF	RC	LAR	ARC	LP	SFF
Equal error rate (%)	0.0604	0.11	0.16	1.01	1.12	2.06	2.19	2.12

621.3
RAJ/F
P07

THESIS
Lakshminath Bezbaroa Central Library
Indian Institute of Technology Guwahati
ACC. No. 1853
28/3/19.