SPECTRAL ANALYSIS OF STRESSED SPEECH FOR SPEECH RECOGNITION

A

Thesis submitted

for the award of the degree of

DOCTOR OF PHILOSOPHY

By

SUMITRA SHUKLA

DEPARTMENT OF ELECTRONICS AND ELECTRICAL ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY GUWAHATI

GUWAHATI - 781039, INDIA

January 2013

Certificate

This is to certify that the thesis entitled "**SPECTRAL ANALYSIS OF STRESSED SPEECH FOR SPEECH RECOGNITION**", submitted by **Sumitra Shukla** (07610203), a research scholar in the *Department of Electronics and Electrical 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 her under our supervision. 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.

Dr. S. Dandapat
Professor
Dept. of Electronics and Electrical Engg.
Indian Institute of Technology Guwahati
Guwahati - 781 039, Assam, India.

Dr. S. R. Mahadeva Prasanna
Professor
Dept. of Electronics and Electrical Engg.
Indian Institute of Technology Guwahati
Guwahati - 781 039, Assam, India.

Dedicated to

My parents

for their love and support

&

My husband

for his love and encouragements

Acknowledgements

I would like to express my gratitude to my supervisors, Professor S. Dandapat and Professor S. R. M. Prasanna for their guidance, help and encouragement throughout my thesis work. I have benefited a lot from them in the area of stressed speech recognition. I am obliged for being introduced to such a relevant research area. I am especially thankful to Professor S. Dandapat for patiently checking all my manuscripts and thesis. Without his enthusiastic support, patience and constant review of the work presented, it is impossible for me to finish this thesis. I also thankful to Professor S. R. M. Prasanna for providing me the financial help during database collection.

I am also thankful to my doctoral committee members Professor P. K. Bora, Dr. R. Sinha and Dr. P. K. Das for sparing their precious time to evaluate the progress of my work.

I would also like to thank the Head of the Department and other faculty members for their help and support in carrying out my work. My special thanks go to my senior Dr. L. N. Sharma for his support during my work and for providing an excellent facility and maintaining a good ambience in the Electro Medical and Speech Technology lab to carry out my work. I would also like to thank Mrs. Jharna, Mr. Sanjib Das and Mrs. Josephine for their timely help.

I would like to acknowledge Dr. P. Krishnamurthy and Dr. Senthil Raja for their suggestions during our discussions and providing a preliminary framework for my research. I would also like to thank my other senior members of EMST Lab Dr. Manikandan, Dr. Nirmala, Dr. Jayanna and Dr. Shweta for their overwhelming support during my PhD course. I would like to thanks my brother Sumit and Abhinav for helping me during stressed speech database collection.

I would like to thank my friends Dr. Debadatta Pati, Mr. Govind and Mr. Haris B. C. for their constant help and enthusiastic company. I highly appreciate the support of my friends Bandita, Padam Priyal, Deepak, Biswajit, Nagaraj, Ramesh, Syed, Rohan, Bandita, Aniruddh, Malaya, Bhanu Priya, Anurag, Sibasankar and Jiss for my work.

I am thankful to my friends Arpana, Minaxi, Monikonkana, Jashmini, and Deepika for all the love and support rendered to me during the course of my stay in IIT Guwahati. I would

like to give my special thank to my friends Swati, Anasuya, Namita, and Bharati for their love, help and continuous motivation. I also grateful to my friend Mrs. Parizat Barua, her husband Dr. Utpal Barua and their daughters mau and mee for making me feel at home away from home.

I thank my parents and my parents-in-laws for their affection, advice and moral support in my life. I also thank my sister Suchitra and my brothers Amit and Sumit for always standing by me and loving me selflessly. I would also like to thanks my sister-in-laws Samridhi, Meena and Hemlata, brother-in-laws Arvind and Manoj, nieces Shivani and Shivangi, and nephews Akash, Mrinal and Ashwin for their love, affection and support.

Last but not the least, I am highly indebted to my husband for all the sacrifices he made for my better future and giving me freedom to take my own decisions. My PhD. endeavor could not be completed without his love and support.

Sumitra Shukla

Abstract

The objective of this thesis is to analyze the stress information in the spectral features of stressed speech. The analysis of stress is focused in the frequency domain, with specific emphasis on various sub-areas in representing this structure-spectrum, subband, and cepstrum. The investigation of stress information includes recognition of speech under stressed condition. In this thesis, four problems of stressed speech recognition are dealt. The first problem deals with the development and evaluation of a stressed speech database. The stress and speech information present in the database are validated by evaluating the stress class and speech information present in the utterances. The stress and speech information are evaluated perceptually as well as by using automatic methods for stress classification and speech recognition, respectively.

Under stressed condition, migration of spectral energy takes place from the lower frequency to the higher frequency. The migration of spectral energy effects the spectral tilt and the subband energy of the speech signal. This has been reported in the literature. Compared to the source, the formants are less affected due to stress. As a part of the second problem, this has been revisited. The conventional method for computation of spectral tilt captures the gross spectral energy information of the speech signal. In the present work, relative formant peak displacement (RFD) is proposed to quantify this variation in formant peaks. The RFD values of second, third and fourth formant peaks are computed as relative displacements of these formant peaks from the first formant peak. A stress classifier is developed to investigate the stress information in the RFD feature.

The migration of spectral energies effects not only the features of the spectrum, but also the energies of subbands. Here, the subband energies are the output of filterbanks placed in the spectrum [1]. Therefore, the third problem deals with the analysis of subband energy of the speech signal. The statistical characteristics such as mean, variance and divergence of subband energy are investigated for different stressed conditions. Mean energy values of subbands for stressed speech are observed to be deviated from that for neutral speech. This deviation is more at higher subbands. The difference energy are proposed across subbands using backward difference of consecutive subbands. It is observed that difference energy of subband is less effected due to stress compared to the other statistical characteristics of the subband energy. Three stress compensation techniques are proposed based on statistical and difference energy based analyses. These techniques are weighted mean, smoothed KLD, and differences of subband energies. The effectiveness of these techniques is evaluated by incorporating these techniques in speech recognition application. From this study, it is observed that these stress compensation techniques require additional stress class information during the recognition, which increases the complexity of the system. Therefore, the performance of speech recognition is dependent on the classification rate of stress classifier. Hence, the fourth problem deals with the development of a stress compensation technique which does not require the stress class information. A subspace projection based approach is proposed to separate the speech and stress information from the stressed speech signal. In this approach, the speech and stress information are assumed to be orthogonal. Orthogonality assumption between speech and stress information is verified experimentally. The contributions of this thesis are as follows:

1. A stressed speech database is developed.
2. Relative formant peak displacement is proposed as a measure for spectral tilt and it is proposed for stress classification.

3. Difference energy is evaluated for stress compensation.
4. A subspace projection based approach is proposed and evaluated for stress compensation.

Keywords: Stressed speech recognition, relative formant peak displacement, difference energy of subband, subspace projection, orthogonality, stress classifier and stress compensation.

Synopsis

The speech recognition system recognizes the spoken words present in an utterance [2]. In the development of conventional speech recognition system, it is assumed that the speech used for recognition is collected from a controlled environment where there is no background noise and reverberation, and also the speakers are not under stressed condition. However, in practice, the environment is uncontrolled in nature [2]. The speech is termed as *Stressed Speech*, if the predominant factor is stress [3]. Stress is defined as the psychological state of the speaker induced due to the emotional, physiological and external reasons [3]. The stress leads to the change in the dynamics of the vocal tract and the excitation source. As a result, the speech characteristics vary from the neutral condition. The features extracted from stressed speech vary from that of neutral speech and this variation leads to the degradation in the recognition performance. State-of-the-art features for speech recognition such as mel frequency cepstral coefficients ($MFCC$) and linear prediction cepstral coefficients ($LPCC$) capture spectral information of the speech signal [4]. The spectral features of the speech signal include formant location, formant peaks, spectral slope, spectral shape and auditory spectral density [5] [6]. This thesis documents the analysis of stress information in three levels, namely, spectrum, subband and cepstral levels. This thesis also proposes methods to model these information and uses these information for the development of robust speech recognition system.

There are four major contributions made in this work. The first one is the development and evaluation of a stressed speech database. First, stressed speech database is developed to analyze the stress information in speech features and the stress

information are evaluated to make this database usable for speech recognition under stressed condition. The migration of spectral energy takes place when speech is stressed. Due to this migration, the spectral features vary as compared to the neutral condition. In the second, the stress information is analyzed on spectrum using gross spectral tilt and local spectral tilt which is termed as relative formant peak displacement (RFD). The stress information in RFD is evaluated using stress classifier where RFD and gross spectral tilt are together considered as feature for stress classification. The migration of spectral energy effects the subband energies. As a third contribution, the stress information is analyzed on subbands using backward differences of subband energy as a feature. The fourth contribution deals with the analysis of stress in cepstral feature using subspace projection approach. A subspace projection based compensation technique is developed for the speech recognition which does not require explicit knowledge of stress classes.

Motivation of the Work

There were four approaches explored in the literature to analyze the stress information in the speech features and use these information to increase the robustness of recognition system. These approaches are multistyle training, classification of stressed speech, robust feature extraction technique and stress compensation technique. In the multistyle training, the speech recognition system is trained using features from the different stressed conditions that may occur during testing [7]. The objective of stress classifier is to recognize the type of stress present in the speech signal and it is used as a prior knowledge for speech recognizer [8] [9] [3]. Robust feature extraction approach was explored to extract features which are immune to the stress and more sensitive to the message present in speech [10]. Ghazale and Hansen [10] observed shifts in spectral energies, and have shown that the highest recognition rate occurred near second formant region rather than first formant region. In this study, the mel filterbank is modified such that the maximum resolution is given near second formant region. In stress compensation approach, the

stressed speech feature vectors are transformed to neutral speech feature vectors. Therefore, neutral speech models can be used for recognition. The adaptive cepstral mean normalization [11, 12] and formant location and bandwidth based compensation techniques were proposed.

The speech recognition system mostly uses cepstral features such as $MFCC$ and $LPCC$. Under stressed condition, the means of cepstral coefficients are shifted from those under the neutral condition [11]. The shift in the means of cepstral coefficients are observed due to the variation in spectral tilt. It is observed that, the migration of spectral energy and variation in vocal loudness take place as a consequence of variations in the glottal parameters [13]. The variation in migration of spectral energy introduces shift in the spectral tilt and variations in spectral energies. The spectral energy contains phonetically relevant information of speech whereas, spectral tilt contains information about speaker such as stress [6] [14]. Alternatively, the variation in vocal loudness is reflected from the formant peaks which also contains stress specific information [15]. These observations inferred that the stress information can be analyzed from spectrum, subband energy and cepstral features. A detail analysis of spectral and cepstral features under stressed condition may provide useful information in the development of robust speech recognizer. The performances of existing approaches for stressed speech recognition are evaluated using speech under simulated and actual stress condition ($SUSAS$) database. This database has less number of utterances for training and testing which may not give the effectiveness of the proposed techniques. Hence, there is need for the development of a stressed speech database.

Stressed Speech Database: Development and Evaluation

In this study, the development and evaluation of a stressed speech database are presented. A simulated stressed speech database is collected in Hindi, an Indian language from fifteen non-professional speakers. A database of one hundred nineteen words is recorded for neutral, angry, sad, Lombard conditions in two separate

sessions. The database is evaluated to understand the level of stress present in utterance and ability of listeners and automatic stress classifier to classify the utterances according to the stress classes. The perceptual validity of this database is evaluated in the similar way in which the $SUSAS$ database is evaluated [16]. The confusion patterns of listeners and an automatic stress classifier are observed to be similar. The listeners are able to identify 63.10% stress classes (in average sense) and the $MFCC$ based stress classifier using VQ and HMM classifiers identify 60.81% and 59.43% stress classes, respectively. The perceptual validity of stress for $SUSAS$ database is evaluated in [16]. Listeners are able to identify angry, loud, neutral and slow speech approximately 52.83%. The $MFCC$ based stress classifier identify approximately 63.25% stress classes. This study infers that perceptually and automatic stress classifier perform approximately similar stress classification in both of the databases. Hence, simulated stressed speech database can be useful for evaluation of speech recognition under stressed condition. The content of utterance of this database is evaluated by listeners and automatic speech recognition method. It is observed that the human perception is not affected by stress condition. For automatic speech recognition, performance of the system under stressed condition is not at par with its own performance when speech is produced under neutral condition.

Analysis of Relative Displacement of Formant Peak

This study focuses on analysis of stress on the spectrum using spectral tilt as feature. Spectral tilt can be defined as the relative distribution of spectral energies from lower frequency to higher frequency [17]. A linear regression line is fitted to the spectrum using least square error method [17, 18]. The slope of the regression line is taken as a measure for spectral tilt. This spectral tilt gives gross energy variation across spectrum, therefore, it is named as gross spectral tilt. The gross spectral tilt is analyzed for different stressed conditions and it is observed that the gross spectral tilt varies under stress conditions. This study infers that gross spec-

tral tilt has ability to characterize the stress classes. This study also inferred that the variation in the spectral tilt is not only due to migration of spectral energy, but also due to the variation in formant peaks that is introduced as a consequence of migration of spectral energy as well as vocal loudness. The effect of stress is also studied on the formant peaks by proposing local spectral tilt as a measure. Here, local spectral tilt refers to the relative variation between formant peaks and it is named as Relative Formant Peak Displacement (RFD) given as

$$RFD_i = \frac{A_1 - A_i}{|A_1|} \qquad 2 \leq i \leq 4 \tag{1}$$

where, A_i is the log magnitude at i^{th} formant location. The RFD values are computed from the LP spectrum and cepstrally smoothed spectrum. Under stressed condition, the RFD features show significant variation and found that the effect of stress is more at higher formant region. The stress information in this feature is evaluated by considering classification of stressed speech. Conventional $MFCC$ feature contains spectral energy information. Both information are extracted from the spectrum. In order to investigate the stress information in RFD feature, the performance of this feature is compared with the performance of $MFCC$ feature. The performance of stress classification using $MFCC$ feature is 53.15%. Alternatively, the performances of RFD features derived from LP spectrum and cepstrally smoothed spectrum are 51.67% and 52.40%, respectively. Results show that RFD feature has approximately same discrimination capability for stress as $MFCC$. $MFCC$ feature captures spectral energies of the speech signal where formant peak information are normalized. On the other hand, RFD feature captures formant peak information. The analysis shows that the information present in the $MFCC$ and RFD are different and hence, they may capture different aspects of stress information. The $MFCC$ and RFD features are combined at feature level, at score level and at rank level. The stress classification performances of all these combination techniques

show improvement from their individual feature performances. From these studies, it can be concluded that the RFD feature contains additional information of stress which is not present in the $MFCC$ feature.

Evaluation of Subband Energies

The migration of spectral energies not only effects the features of the spectrum, but also energies of subbands where the subband energies are the output of filterbanks placed in the spectrum [1]. This study inferred that the statistical characteristics of subband energies (SBE) vary under stressed conditions. Thus, statistical parameters, namely, mean and Kullback Leibler divergence (KLD) of SBE are considered to investigate the stress information in individual subband. Analyses show that the means of SBE under angry and Lombard conditions are shifted to the higher level and means of SBE of sad speech is shifted to the lower level. Further, higher subbands are deviated more from that of neutral speech than lower subbands. From this analysis, it is also observed that the variation patterns of means of energies across subbands are observed to be same under stressed conditions. In this study, this variation patterns of SBE is estimated using backward difference between consecutive subbands. This may contain the dynamic information of the subbands which might give the information of rate of migration of energies across frequency scale. In this study, the backward differences of subband energies is proposed ($BDSBE$) and it is given in Eq. 5.2.

$$\Delta E^X(k) = E^X(k) - E^X(k-1) \tag{2}$$

where, k: Subband number, E: Subband energies, ΔE^X: Differences of subband energies ($DSBE$) of X stress class. It is observed that up to 10 subbands, $DSBE$ of stressed speech vary from neutral speech and after 10^{th} subbands, the variation is approximately same for stressed conditions. In mel-scale, 10 subbands correspond to the 1000 Hz (approximately) frequency region which may contain 1^{st} and 2^{nd}

formant information. This observation infers that the migration of energy from 1^{st} formant to the 2^{nd} formant introduces modulation in higher formants due to stress. Therefore, the energies of higher subbands under stressed condition are more deviated from those under neutral condition. From this study, it is also observed that the variation of means of $DSBE$ are negligible across stressed conditions compared to those of means of SBE. This study inferred that the $DSBE$ may contain stress robust information of the speech signal. In order to investigate the effectiveness of these analyses, stress compensation technique is proposed. The objective of this compensation technique is to eliminate the effect of stress from the subbands. In this study, compensation techniques are proposed to transform the energies of subbands i.e. $E^X(k)$ of given stress X to energies of subbands $\hat{E}^X(k)$ which is approximately close to neutral speech using weighting factor $W^X(k)$ as given in Eq. 5.3.

$$\hat{E}^X(k) = W^X(k)E^X(k) \tag{3}$$

From analysis of means of SBE under different stress conditions, weighted mean based stressed compensation technique is proposed

$$W_1^X(k) = \frac{\frac{1}{M_1}\sum_{i=1}^{M_1} E_k^N(i)}{\frac{1}{M_2}\sum_{i=1}^{M_2} E_k^X(i)} \tag{4}$$

where, X: Stress condition, E_k^X: Energy of k^{th} subband for stress X, M_1 and M_2: Number of frames for neutral and stress class, respectively, W_1^X: Weighting factor for proposed mean normalization of SBE for a given stress. KLD information of subband energies is also used to compensate the effect of stress from the speech

signal as given below

$$W_2^X(k) = \begin{cases} (1 + \frac{D_k}{max(D_k)}) & \text{for sad speech} \\ \frac{1}{(1+\frac{D_k}{max(D_k)})} & \text{for angry and Lombard speech} \end{cases} \quad (5)$$

where, D_k: Divergence value of stress speech from neutral speech of k^{th} subband. The $DSBE$ is used as mapping function for SBE of stressed speech to neutral speech. The recursive approach is proposed to map the subbands of stressed speech to neutral speech.

1. Initialization

$$\hat{\mu}_E^X(1) = \mu_E^X(1) \quad (6)$$

2. Recursive

$$\hat{\mu}_E^X(k+1) = \Delta E^N(k) + \hat{\mu}_E^X(k) \quad (7)$$

3. Weighting factor

$$W_3^X(k) = \frac{\hat{\mu}_E^X(k)}{\mu_E^X(k)} \quad (8)$$

The effectiveness of compensation techniques is evaluated for stressed speech recognition. The performances of these proposed compensation techniques are evaluated under the assumption that the recognition system has prior knowledge of stress classes and their performances are compared with $MFCC$ feature based speech recognition. The performances of weighted mean, smoothed KLD and $DSBE$ based compensation techniques are 64.89%, 63.19%, and 65.31%, respectively, which are 5.07%, 3.37% and 5.49%, respectively higher than $MFCC$ feature. However, the stress information is not known to the recognizer in a practical scenario. The performances of these compensation techniques are evaluated using rank level combination of $MFCC$ with RFD features. The performances of weighted mean, smoothed KLD and $DSBE$ based compensation techniques are 62.63%, 61.14%, and 63.34%, respectively. These results infer that the performance of the recognizer depends on

the classification of the stress classifier. Any failure of the stress classifier reduces the performance of the system.

Stress Analysis using Subspace Projection

This chapter proposes a stress compensation technique which does not require the explicit knowledge of the stress class. The subspace projection based technique is proposed to separate the speech information from stressed speech signal. In this technique, an orthogonal relation is assumed between speech and stress components. The projections of stressed speech vectors onto the neutral speech subspace form speech subspace and the stress subspace which is orthogonal to the neutral speech subspace. In this study, the codevectors of the speech, $\mathbf{A} = \{\mathbf{a}_1, \mathbf{a}_2, ..., \mathbf{a}_N\}$ are assumed as neutral speech vectors. Under stressed condition, the speech vector contains additional stress information. This vector can be termed as stressed speech vector. The stressed speech vector ($\mathbf{y}$) can be represented as a function (f) of speech ($\mathbf{s}$) and stress ($\mathbf{x}$) information. Under stressed condition, the stressed speech vector deviates from neutral speech vector. In this study, this deviation is assumed due to stress. Let, $\mathbf{P}_n$ be the projection matrix required to project a vector onto $\mathbf{a}_n$. Then $\mathbf{P}_n$ is given by

$$\mathbf{P}_n = \frac{\mathbf{a}_n \mathbf{a}_n^T}{\mathbf{a}_n^T \mathbf{a}_n} \qquad 1 \leq n \leq N \tag{9}$$

The projection of $\mathbf{y}$ onto $\{\mathbf{a}_n\}$ which can be represented as $\mathbf{p} = \{\mathbf{P}_n \mathbf{y}\}$ where, $\mathbf{p}$ is the component of $\mathbf{y}$ in the neutral speech subspace. The set of projections of stressed speech vector, $\mathbf{y}$, onto the set of codevectors $\{\mathbf{a}_n\}$, $1 \leq n \leq N$, can be seen as projection of $\mathbf{y}$ onto neutral speech subspace. These components contain speech specific information of stressed speech vector. The error between $\mathbf{y}$ and $\mathbf{P}_n \mathbf{y}$ is orthogonal to the neutral speech subspace which can be represented as $\mathbf{e} = \mathbf{y} - \mathbf{P}_n \mathbf{y}$. Due to orthogonal assumption between speech and stress information, the orthogonal vectors will contain stress information of stressed speech vectors.

The projection of $\mathbf{y}$ onto the set of $\{\mathbf{a}_n\}$, produces N projected vectors in neutral

speech subspace and corresponding N orthogonal vectors in the orthogonal subspace. The stressed speech subspace can be decomposed into the speech ($\mathbf{S}$) and the stress ($\mathbf{X}$) subspaces. In this work, the stress vector $\mathbf{x}$ is decided by the orthogonal vector which has minimum length as given in Eq. 6.6 and Eq. 6.7. The projected vector corresponding to that orthogonal vector is considered as speech vector as given in Eq.6.8.

$$\hat{n} = \arg\min_{n}[(\mathbf{y} - \mathbf{P}_n\mathbf{y})^T(\mathbf{y} - \mathbf{P}_n\mathbf{y})] \tag{10}$$

$$\mathbf{x} = (\mathbf{y} - \mathbf{P}_{\hat{n}}\mathbf{y}) \tag{11}$$

$$\mathbf{s} = \mathbf{P}_{\hat{n}}\mathbf{y} \tag{12}$$

According to orthogonal assumption, the speech information should be present in speech subspace and stress subspace should contain stress information. The speech information in stress subspace ($\mathbf{X}$) should be negligible and similarly little stress information may be present in speech subspace ($\mathbf{S}$). This assumption is verified experimentally by using speech and stress recognition techniques. The performances of speech recognition using stressed speech vectors and their corresponding estimated speech and stress vectors are evaluated. This study show that the recognition performances of estimated speech vectors under stressed conditions are increased from those of stressed speech vectors. The speech recognition performances of estimated stress vectors under stressed conditions are less than those of stressed speech vectors and estimated speech vectors. Alternatively, the classification rate of stress classifier is evaluated using stressed speech vectors and their corresponding estimated speech and stress vectors. This study show that the stress classification rates of estimated speech vectors under stressed conditions are less than those of estimated stress vectors and the classification rates of estimated stress vectors under stressed conditions are increased from those of estimated speech vectors. The speech and stress recog-

nition experiments concluded that the speech and stress specific information are present in their respective subspaces.

Table 1: Performances of proposed techniques for stressed speech recognition

	$MFCC$	Multistyle training	Adaptive CMN	$RFD(log)$ $+MFCC$	Difference Energy	Subspace Approach
Neutral	81.10	80.42	73.54	72.62	76.73	80.96
Angry	42.58	64.82	39.86	69.72	52.11	44.05
Sad	59.89	63.64	53.43	63.99	58.87	63.91
Lombard	55.69	75.14	52.14	75.19	65.63	56.49
Avg. Perform	59.82	71.14	54.74	70.38	63.34	61.35

The estimated speech and stress vectors are used for speech recognition. The performances of proposed techniques developed based on the analysis of stress are evaluated for speech recognition and they are shown in Table 3.10. These performances are compared with the performances of $MFCC$ feature, existing techniques such as multistyle training and adaptive CMN. Different columns indicates different features or techniques of speech recognition under stressed condition. The proposed rank level combination of $MFCC$ and RFD feature is evaluated for stress directed speech recognition and it is named as $RFD(log)$-$MFCC$. By comparing the performances of multistyle training and $RFD(log)$-$MFCC$, it is observed that the performances of $RFD(log)$-$MFCC$ technique under stress conditions improve from the multistyle training. The subspace projection and difference energy based compensation techniques provide better performance compared to adaptive CMN techniques. Also, the proposed subspace projection based compensation technique does not require additional stress information during recognition. Therefore, this technique is computationally easier to deploy for speech recognition under stressed condition.

Contents

List of Figures

List of Tables

List of Acronyms

ASR	Automatic Speech Recognition
ANN	Artificial Neural Networks
ASRFD	Average Stressed Relative Formant Peak Displacement
AC	Autocorrelation of Mel Frequency Cepstral Coefficients
CDCN	Codeword Dependent Cepstral Normalization
CMN	Cepstral Mean Normalization
CDHMM	Continuous Density Hidden Markov Model
DCT	Discrete Cosine Transform
DFT	Discrete Fourier Transform
DTW	Dynamic Time Warping
DSBE	Difference Energy of Subbands
DET	Detection Error Trade-off
EVD	EigenValue Decomposition
FFT	Fast Fourier Transform
HMM	Hidden Markov Models
IDFT	Inverse Discrete Fourier Transform
KLD	Kullback-Leibler Divergence
KNN	K-Nearest Neighbor
LP	Linear Prediction
LPC	Linear Predictive Coefficients
LPCC	Linear Predictive Cepstral Coefficients
LFPC	Log Frequency Power Coefficients

MAP	Maximum *a Posteriori* Adaptation
MFCC	Mel-Frequency Cepstral Coefficients
PDF	Probability Density Function
PLP	Perceptual Linear Prediction Cepstral Coefficients
RMS	Root Mean Square
RFD	Relative Formant Peak Displacement
SUSAS	Speech Under Simulated and Actual Stress
SUSE	Speech Under Simulated Emotion
SUSSC	Speech under Simulated Stress Condition
SRFD	Stress Relative Formant Displacement
SBE	Subband Energy
TEO	Teager Energy Operator
VQ	Vector Quantization

List of Symbols

a	Pre-emphasis filter coefficient
a_k	Linear predictive coefficients
A_i	Peak at i^{th} formant location
$\mathbf{A} = \{\mathbf{a}_n\}$	Vector of neutral speech subspace
$\mathbf{B} = [b_{jk}]$	Observation symbol probability
$c_l(n)$	Cepstral coefficients of speech signal at time n
c_r and c_t	Reference and test cepstral coefficients
C	Covariance matrix
d	Spectral distance
D_{KL}	KL-divergence
$\mathbf{e}$	Orthogonal vector
$e_h(n)$	Hilbert transform of the LP residual
E	Subband energy
$\hat{E}$	Modified subband energy
ΔE	Difference energy
f	Real frequency
f_{mel}	Perceived frequency
F_s	Sampling frequency
$H(Z)$	Vocal-tract system filter transfer function
k	Subband index
K	Number of Subbands
l	Coefficient index

L	Number of Mel cepstral coefficients
λ_w	Word model
M	Number of observations
μ_i	Mean vector
n	Discrete time signal time index
N	Number of vectors
o	Observation sequence
p	Linear prediction order
$\mathbf{P}$	Projection matrix
$\mathbf{p}$	Projected vector
$P(\lambda\|o)$	HMM likelihood
$s(n)$	Speech signal
$S(k)$	Output power of k^{th} subband
$(\mathbf{S})$	Speech subspace
$(\mathbf{s})$	Speech vector
σ_E	Standard deviation of subband energy
σ^2	Gain of LP coefficients
T	Observation length
$\mathbf{T}$	Transition matrix
$\boldsymbol{U}$	Orthogonal matrix
$\mathbf{\Lambda}_y$	Diagonal matrix
W	Weighting factor
w_i	Weights
$\mathbf{x}$	Stress vector
$\mathbf{x}$	Stress subspace
X	Stress class
$\mathbf{y}$	Stressed speech vector
$\mathbf{Y}$	Stressed speech subspace

1

Introduction

Contents

1. Introduction

The speech recognition system recognizes the spoken words present in an utterance [2]. In the development of conventional speech recognition system, it is assumed that the speech used for recognition is collected from a controlled environment where there is no background noise and reverberation, and also the speakers are not under stressed condition. However, in practice, the environment is uncontrolled in nature [2]. The speech is termed as *Stressed Speech*, if the predominant factor is stress [3]. Stress is defined as the psychological state of the speaker induced due to the emotional, physiological and external reasons [3]. The stress leads to the change in the dynamics of the vocal tract and the excitation source. As a result, the speech characteristics vary from the neutral condition. The features extracted from stressed speech vary from that of neutral speech and this variation leads to the degradation in the recognition performance. Four stages are considered in stressed speech processing to improve the robustness of speech recognition for stressed speech [3]. These are stress analysis, stress classification, stress compensation and robust speech recognition. In stress analysis stage, the stress information is investigated in speech features for identifying their robustness or sensitiveness towards stress. In stress classification stage, stress class is identified using stress sensitive speech features. In stress compensation stage, the effect of stress is removed from the speech signal using stress robust features. Finally, in robust speech recognition stage, the robustness of speech recognition is increased under stressed condition by exploring new feature for recognition or by incorporating stress classification or stress compensation techniques in the speech recognition system. Exploring each of these stages is the objective of the present work.

The first objective of the work will be to develop and evaluate a stressed speech database. The second objective of the work will be to analyze the effect of the stress information at different levels, namely, spectrum, subbands and cepstral levels. The third objective of the work will be to develop methods to model the stress information, that can be used in stress classification or in stress compensation. Finally, the fourth objective of the work will be to incorporate these modeling techniques into the speech recognition system.

1.1 Overview of Automatic Speech Recognition System

Speech is the primary mode of conveying message in human-human communication. An automatic speech recognition (ASR) system is developed to use speech as a mode for human-machine communication. The ASR system recognizes the spoken words present in an utterance [2]. Speech recognition is deployed in several areas such as commercial mobile communication, where it handles voice dialing, command and control, voice enabled short messaging service (SMS), email, and mobile search [19]. In military services, it handles several tasks like setting radio frequencies, commanding an autopilot system, setting steerpoint coordinates and weapons release parameters [20–22]. There are some other areas where speech recognition is used such as voice enabled form filling, personal data assistants, car navigation, online ticketing systems, customer care management systems for call center purposes and development of vocoder in speech communication [23].

The speech production system consists of vocal cords and vocal tract system [2]. The vocal tract system consists of various articulators such as jaws, tongue, velum, lips and mouth. From the speech production point of view, speech is produced as a sequence of sounds that depends on variation of the state of vocal cords as well as shapes, positions and sizes of various articulators with respect to time [2]. Speech can be classified as voiced speech and unvoiced speech depending on the state of vocal cords [2, 24]. Speech is called as a voiced speech, when the vocal cords are tensed and vibrate periodically, as air flows from the lungs. The nature of voiced speech is quasi-periodic. Alternatively, the unvoiced speech is produced when the vocal cords are relaxed and not vibrating when air flows from the lungs. The resulting speech is aperiodic in nature. From the signal point of view, speech is a time varying signal whose characteristics are approximatively stationary over short period of time (10-30 msec). The characteristics of the signal change over long period of time ($\geq$ 100 msec) which reflect the different speech sounds being spoken. The speech signal can be characterized via spectral representation of the model of speech production system. Human vocal tract is a concatenation of the vocal tubes, of varying cross-sectional areas, that is excited by the vocal cords. According

to the acoustic theory, the transfer function of energy from the excitation source to the output produces resonances of the vocal tube, where the most of the acoustic energy passes from the source to the output [2]. These resonances are called formants of speech. In the speech signal, five formants are present which characterize the content of the speech signal.

In speech recognition research, several issues have been addressed to make speech recognition system more sophisticated, intelligible and robust to the environmental conditions. Initially, the research was more focused toward exploiting the fundamental ideas of acoustic-phonetic [25] [26]. Speech specific information is mainly captured by the vocal tract parameters such as formant locations and their bandwidths, spectral shapes and auditory spectral density. Therefore, the research was carried out to extract features which capture the vocal tract parameters more accurately. These features are formant locations and their bandwidths, filterbank based mel frequency cepstral coefficients ($MFCC$) [4], linear prediction cepstral coefficients ($LPCC$) [27], and perceptual linear prediction cepstral coefficients ($PLPCC$) [28]. The speech recognition using these features perform well for a single speaker and for a small vocabulary [29]. However, the acoustic properties of phonetic units are highly variable for speakers and neighboring phonetic units due to the coarticulation of sounds. To overcome these limitations, the pattern classification approach was explored for speech recognition [29] [30]. In this approach, the speech knowledge is brought into the system via the training procedure, where the machine learns which acoustic properties of the speech class are reliable and repeatable across all training patterns. To develop a speaker independent system and to handle a large vocabulary size, several pattern classification and pattern matching techniques have been explored such as dynamic time warping (DTW) [31], vector quantization (VQ) [29], hidden Markov model (HMM) [30] [2] and artificial neutral network (ANN) [32].

The speech recognition system is developed with the assumption that it should be deployed in a controlled environment, where the noises and distortions are negligible [33]. However, the performance of speech recognition rapidly degrades in presence of noises and distortions [34] [35] and other mismatches between training and testing conditions, which include microphone mismatch [36], transmission channel mismatch. Some noises are additive in nature. While record-

ing, noises are added with the speech signal. Mostly, these types of noises are ambient noises such as in office environment. The source of noises include typewriter, printer, telephone ringing, sound of fans and background conversation of other people. Some noises are convoluted with the speech such as reverberation [37]. The microphone transducer, depending on its type and mounting position, also introduces distortion in the speech spectrum [36]. Several normalization techniques were explored to enhance the speech signal characteristics such as parallel model composition [35], model decomposition [34], and spectral subtraction methods. These techniques were used for speech enhancement and noisy speech recognition. Codeword-dependent cepstral normalization ($CDCN$) technique was proposed for microphone mismatching [36]. The development of speaker independent continuous speech recognition was also considered as serious problem in the research [38]. The speaker variability introduces variation in the feature which is difficult to handle with statistical models. Therefore, speaker independent models are less effective compared to the speaker dependent models for speech recognition. Further, the development of large number of speaker model is also computationally expensive. To handle this issue, the speech recognizer was developed using small number of speakers and new speaker's speech was used for tuning the model for new speakers. This technique is termed as speaker adaptation. Recent speech recognition research is mostly focused on the robustness from the natural or unconstrained human-human speech of the broadcasts [39], telephone conversational speech [40] and spontaneous speech [41, 42].

The general block diagram of automatic speech recognition system is shown in Figure 1.1. Speech recognition consists of two phases, training and testing. In the training phase, the recognizer is able to characterize the acoustic properties of the speech pattern that can be sound, word or phrase by using adequate training set of that pattern. On the other hand, in the testing phase, the unknown patterns are compared with each possible pattern learned during the training phase and classifies the unknown patterns according to the goodness of match of the patterns. The training phase includes feature extraction and modeling blocks. The testing phase includes feature extraction, pattern similarity and decision.

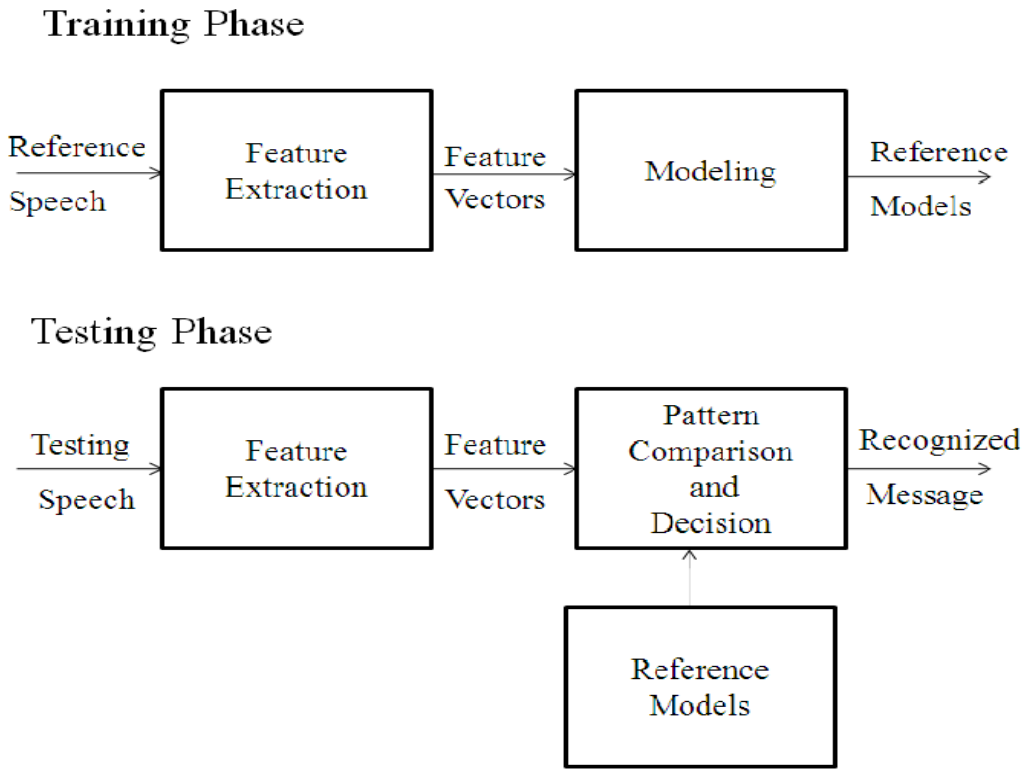

Figure 1.1: Block diagram of automatic speech recognition system

1.1.1 Feature Extraction

Feature is the compact representation of the acoustic properties manifested in the speech signal [2]. The speech recognition system uses features which represent the vocal tract information adequately. Various signal processing techniques have been explored to extract the vocal tract characteristics more effectively.

Davis et al. [25] considered formant frequencies for a speaker dependent isolated digit recognition. The formant frequencies are distinct for different sounds, therefore, formant frequencies were measured for vowel regions and the trajectories of the first and second formant frequencies were plotted in a two dimensional space. The formant trajectories were observed to be distinct for each digit. Forgie and Forgie [26] developed speech recognizer to recognize vowels. In this study, spectral patterns of steady vowel region were determined using filterbank energy con-

cept. The filterbank energy is influenced by variations in the pitch frequency. To tackle this, Ichikawa and Nakata [43] introduced linear prediction (LP) analysis in the speech recognition field. The LP analysis estimates linear, time varying system which is closely matched with the speech production model. Here, vocal tract system is characterized by linear and time-varying system [2]. Davis and Mermelstein [4] evaluated $MFCC$, linear frequency cepstrum, linear prediction cepstrum, a linear prediction spectrum, and a set of reflection coefficients features for large vocabulary continuous speech recognition. It was observed that the $MFCC$ feature performs better than other features. $MFCC$ is a perceptually motivated feature, which gives linear frequency resolution up to 1000 Hz and logarithmic resolution at higher frequencies [2]. As a result, it suppresses the insignificant spectral variations in the higher frequency bands and provides a better representation of lower frequency bands, which is important from the speech recognition point of view. Furui [44] proposed combination of instantaneous and dynamic features of the speech spectrum for isolated word recognition. It was found that, the concatenation of static and dynamic features perform better for speaker independent scenario. Hermansky [28] estimated auditory spectrum using three concepts of psychophysics of hearing: (1) the critical-band spectral resolution, (2) the equal loudness curve, and (3) the intensity-loudness power law. In comparison to LP spectrum, this technique is more effective to suppress the speaker dependent details from the auditory spectrum and also more consistent with human hearing. For convolution noise, the cepstral coefficients computed from this auditory spectrum derived from perceptual linear prediction analysis gives comparable performance to the conventional spectral subtraction method [45].

1.1.2 Modeling

The pattern recognition approach has two stages namely, training of speech patterns and recognition of speech patterns via pattern comparison [2]. In pattern training, one or more patterns corresponding to speech sounds of the same class are used to create a pattern, which is a representative of that class. Here, pattern refers to the feature vectors of sound. The resulting pattern is called a reference pattern. It can be derived through some averaging techniques or it

can be a model that characterizes the statistical characteristics of the reference pattern. The modeling techniques are K-nearest neighbor (KNN) decision, vector quantization (VQ) [46], hidden Markov model (HMM) and artificial neural network (ANN) [32].

Rosenberg and Soong [46] introduced VQ technique in speech recognition area, to model the reference patterns of words. In the VQ technique, acoustically similar features formed clusters and the centroid of each cluster is termed as a codevector. Lippmann [32] introduced self-organizing ANN technique in the speech recognition field. Neural network basically consists of neurons that are connected via weights that are adapted according to the speech specific information during training. Rabiner [30] applied stochastic process based HMM for speaker independent isolated word recognition. In this experiment, multiple codebook based HMM i.e. discrete HMM and continuous HMM, where observation sequences are assumed to be Gaussian distributed, were developed. It was observed that continuous HMM models the feature vectors more accurately than discrete HMM. The discrete and continuous HMM methods were found to be suitable for handling speaker independent isolated word recognition [2].

1.1.3 Pattern Comparison and Decision

In pattern comparison approach, the unknown test pattern is compared with the reference pattern of each class and the similarity is measured between the test pattern and each of the reference patterns. Usually, the test pattern and reference pattern of same sound units (word/ sentence) may not have same duration. The fluctuation in duration may be due to variation in the speaking rate. To eliminate the nonlinear durational fluctuation between these two patterns and to measure the similarity between two patterns, dynamic time warping (DTW) was developed by Vintsyuk [47]. In this technique, the duration of unknown patterns was normalized with respect to the reference pattern. Element-by-element recognition takes place ensuing best match between elements of unknown and reference patterns. Weighted cepstral distance based similarity measure has been proposed in speech recognition area. The weighted

cepstral distance from test pattern, c_t to the reference pattern, c_r can be represented as

$$d(c_t, c_r) = (c_t - c_r)\mathbf{C}^{-1}(c_t - c_r)^T \tag{1.1}$$

where, $\mathbf{C}$ is positive definite matrix. The Euclidean distance is widely used as a distance measure technique for VQ technique. For Euclidean distance measure, $\mathbf{C}$ is considered as an identity matrix, whereas, for Mahalanobis distance, $\mathbf{C}$ is considered as a covariance matrix of the feature vectors [48]. For LP coefficients based spectral feature, Itakura-Saito distance was proposed [44,49]. For large vocabulary speech recognition system, probabilistic model of speech production is used [50]. Using Bayes rule, the probability that an observation sequence o was produced by the model λ_w is given as

$$P(\lambda_w|o) = \frac{P(o|\lambda_w)P(\lambda_w)}{P(o)} \tag{1.2}$$

where, $P(o|(\lambda_w)P(\lambda_w)$ is the probability of a sequence of observations from the model λ_w. $P(\lambda_w)$ is the probability associated with a postulated sequence of word model.

The decision stage decides which reference model matches most closely with the unknown test speech [4]. For a VQ based classifier, the average Euclidean distance of test speech with the closest reference model must be minimum compared to the other reference models. Similarly, for probabilistic model, the maximum a posteriori probability (MAP) decoding rule is used for decision. The average $P(\lambda_w|o)$ of test speech with the closest reference model must be maximum compared to the other reference models, which is given in Eq. 1.3.

$$\hat{w} = \underset{1 \leq w \leq W}{\arg\max}[P(\lambda_w|o)] \tag{1.3}$$

The closest reference model is then finally decided as the basic unit information (message) present in the testing feature vectors.

1.2 Speech Recognition under Stressed Condition

In the development of conventional speech recognition system, it is assumed that the speech used for recognition is collected from a controlled environment. Controlled environment refers to the case where there is no background noise and reverberation, there is no speech from other speakers and also the speakers are not under stressed condition. However, in practice, the environment is uncontrolled in nature, where the speech is affected by one or more of the above mentioned factors. The speech is termed as *Stressed Speech*, if the predominant factor is stress [3]. Stress is defined as the psychological state of the speaker induced due to the internal factor, external factor or both [3]. The emotional and physiological aspects represent the internal factor. The surrounding environmental aspect represents the external factor. Emotionally induced stress includes those due to anger, happiness, fear, anxiety, sorrow and compassion. Physiologically induced stress includes those due to sickness, cough and drug interaction. Externally induced stress includes those due to Lombard, fastness, slowness, question, loudness and softness. The speaker is said to be stressed, if the speaker is affected by one or more of these mentioned conditions. Otherwise, the speaker is said to be normal or neutral. Stress leads to variation in the constriction of speech production organs, which in turn changes the dynamics of the vocal tract and the excitation source. Further, the variation may also be unique for each of the stressed condition resulting in unique variation of speech characteristics from the so called neutral condition. Therefore, the features extracted from stressed speech are different as compared to the features from neutral speech and this variation leads to the degradation in the performance.

Lippmann et al. [51] demonstrated the effect of stress on speech recognition system. A speaker dependent isolated word recognition was developed, which gave approximately 92% recognition performance under neutral condition [51]. However, the performance of the same system degraded to 61% under noisy and stressed condition. This study inferred that the stress information influences the characteristics of speech, which may also interfere with the speech information. As a result, the recognition performance degrades. A detailed investigation of

stressed speech may provide information of stress, expected to be useful for the development of a robust speech recognition system.

1.2.1 Stages Involved in Stressed Speech Recognition

Stressed speech recognition deals with the recognition of speech produced under stressed condition. The objective of stressed speech recognition system is to provide knowledge of stress specific information to the speech recognition system for improving the performance under stressed condition. In general, four broad stages are considered in stressed speech recognition. Brief descriptions of these tasks are given below.

(i) *Stressed Speech Analysis:* This is the first and fundamental stage in the stressed speech recognition, which deals with the investigation of stress information in speech features. The stress information in speech features is in turn reflected from the changes in the characteristics of different features. Stress discriminating capability of different features is also studied to learn about their robustness or sensitiveness towards stress.

(ii) *Stress Classification:* This stage deals with the identification of stress present in the speech signal. This is useful in two areas: objective stress assessment and improving speech processing tasks [3]. Objective stress assessment detects speaker stress class which may be useful in emergency telephone message sorting and aircraft voice communications monitoring and call centers [52]. Stress classification also gives stress specific information to speech processing applications that can be used to improve the performance of the system.

(iii) *Stress Compensation:* This stage deals with the removal of stress present in the speech signal. This is useful for improving the performance of different speech processing applications. For instance, the features of a particular stressed condition may have unique distribution characteristic. These can be estimated in terms of distribution parameters using training data and can later be used for compensation. Such a compensation may result in a modified distribution, which is closer to the neutral speech.

(iv) *Stressed Speech Recognition:* This stage deals with speech recognition under stressed condition. The objective of this stage is to increase the robustness of speech recognition under stressed condition.

1.2.2 Approaches for Stressed Speech Recognition

Based on the tasks associated with the stressed speech recognition, there are four approaches available in the literature. These approaches are namely, multistyle training, stress classification, robust feature extraction and stress compensation.

1.2.2.1 Multistyle Training

The multistyle training focuses on increasing the robustness with respect to the variability present in the speech during the training. The speech recognition system is trained using features from the different stressed conditions that may occur during the testing. Since the speech recognition system has exposure to the different stressed conditions during the training, it is expected to provide improved performance. Lippmann et al. [51] proposed multi-style training using simulated stressed speech utterances. A speaker dependent speech recognition system of 35 word vocabulary was developed. The system was trained using five types of stresses namely, anger, fastness, slowness, soft and loudness, along with the neutral speech. The speech recognition was developed using $MFCC$ as a feature extraction technique and word models were developed using HMM classifier. The average error rate of neutral trained system was observed to be 20.7%. Alternatively, the average error rate of multi-style trained system was 9.8%. This result shows that the multi-style trained system reduces the error rate by a factor of 2.

In another study, Hansen and Ghazale [53] trained the system using artificially generated stressed speech rather than considering original simulated stressed speech. The stressed speech was generated by exploiting the spectral content (i.e. mel cepstral parameters) and duration of voiced and unvoiced segments of speech. Spectral and duration models for each phoneme under stressed condition were developed, and according to the stressed condition the spectral content and the duration of neutral speech were modified. This artificial stressed speech was

then considered for training of the speech recognition system. In this study, a thirty five word vocabulary speaker independent HMM based speech recognizer was developed. The system was trained using four types of stresses namely, Lombard, loud, slow and neutral. The neutral speech trained system gave 57.33% average recognition rate, whereas, 72.67% average recognition rate was obtained from artificial stressed speech trained system. These results show that 15% improvement in the recognition performance was achieved from artificial stressed speech based multistyle training. The same authors also generated artificial stressed speech using five separate perturbation models, namely, voiced duration variation, pitch contour perturbation, derivative of pitch contour perturbation, explicit state occupancy for pitch perturbation HMM and average spectral mismatch [54]. Each of the statistically generated perturbation model was considered for the modification of neutral speech parameters. In this study, a thirty five word vocabulary speaker independent HMM based speech recognizer was developed using three types of stresses namely, Lombard, loud and neutral. Average recognition rate of artificial stressed speech trained model under stressed condition was 84.34%. Alternatively, the average recognition rate of neutral trained system under stressed condition was 77.31%. The relative improvement of 7.5% in performance was achieved as compared to neutral speech trained model.

1.2.2.2 Recognition Using Stress Classification

The objective of stress classifier is to recognize the type of stress present in the speech signal [55]. This may be useful for improving the performance of speech recognition under stressed condition. The output of stress classification system can be used in two ways for improving the speech recognition system performance [8] [9].

(i) *Stress dependent speech processing*: In this approach, during the training, one model is trained for each basic unit (phoneme, syllable, word) per stress. During the testing, the stress classifier identifies the stress present in the speech signal and the feature of the speech signal is tested using only the trained basic unit models for the corresponding stress. For each stress class, variation among the features is due to the different basic

units and hence, an improved performance is expected during speech recognition.

(ii) *Stress compensation*: During training, basic unit models are trained using neutral speech. During testing, the stress classification system identifies the stress present in the speech signal and the stress information from the speech feature is removed using the corresponding stress compensation. The compensated features of speech are tested with the neutral trained models. After compensation, the compensated features of speech are close to that of the neutral speech and hence an improved performance is expected.

Womack and Hansen [3] developed a speaker and text independent stress classifier. In this study, the stress information were analyzed on vocal tract, excitation and cepstral based features. Only a subset of features which had the capability of distinguishing more than two stressed conditions was considered for stress classification. Articulatory cross-sectional area ratio feature of the vocal tract was found to be sensitive for stress. Pitch, duration and intensity features from excitation and autocorrelation of mel cepstral coefficients and their mean and standard deviation features were considered for stress classification. The set of these features was trained with neural network to estimate the stress score, which measures the degree of stress present in the unknown speech signal. Thirty five word vocabulary and neutral, angry, Lombard, loud, slow, soft and fast stressed conditions were considered for the study. Average classification rate for the stress classifier was 91.01%. This stress classifier was further used in the speech recognition task under the stressed condition. An average recognition rate for all the stressed conditions was found to be 80.6%. In case of neutral trained system, the average recognition rate was 69.5%.

Cairns and Hansen [8] found that nonlinear energy of speech signal is mainly produced due to the creation of vortices, when the glottal air flows from true vocal folds to the false vocal folds. Teager energy operator (TEO) was used to measure the nonlinear energy of the speech signal [8]. In this study, it was found that within a pitch period, the nonlinear energy of the speech signal varies under different stressed conditions. This feature was therefore, considered as a feature for stress classification. The classification rate of neutral speech was

97%, loud speech was 98%, angry speech was 99% and Lombard speech was 86%. Ramamohan and Dandapat [9] found that the frequency, amplitude and phase of the speech signal varies under stressed conditions. The sinusoidal model based features were therefore considered for stress classification. The classification performance was evaluated for one Indian language (Telugu) and English language. The average classification rate of stress for Telugu speech was 92.3% and 89% for frequency and amplitude features, respectively. For English speech, it was 87.1% and 76% for frequency and amplitude features, respectively. Casale et al. [56] proposed genetic algorithm based criteria for selection of a set of features for stress classification. These features are autocorrelation of $MFCC$, fundamental frequency, formant frequencies, log area ratio, log energy, LPC, $LPCC$, LSF, $MFCC$, real cepstrum based coefficients, reflection coefficients, the variance of the Linear Prediction Error, and TEO. These features were used for classification of neutral, angry, loud and Lombard speech. Average classification rate for the stress classifier was 98.56%. Väyrynen et al. [57] proposed multiple KNN classifier based decision level fusion technique for classification of stressed speech. The classification system was developed using Finnish emotional speech database. Average classification rate for the stress classifier was 43.5%. Xiao et al. [58] investigated physical parameters based on a two-mass vocal fold model. This model estimates the stiffness of vocal folds, vocal fold viscosity loss, and subglottal pressure coming from the lungs. This parameter is used for classification of stress.

1.2.2.3 Recognition Using Robust Features

The objective of this approach is to extract the feature which is immune to the stress and sensitive to the message present in the speech. Ghazale and Hansen [10] found the highest recognition rate of neutral speech around the region of first formant frequency. On the other hand, the highest recognition rate was observed around the region of second formant frequency for angry speech. The shift in the recognition performance was observed due to the variation in slope of the spectrum. The existing mel scale gives more resolution up to 1000 Hz and less resolution at frequencies greater than 1000 Hz. Therefore, the feature warped by mel scale can model first formant frequency more correctly as compared to the second formant

frequency. In this study, the mel scale was modified in such a way that it gives more emphasis near 2^{nd} formant region rather than 1^{st} formant region. Two new scales, namely, modified mel scale and expolog scale were proposed, which gave more emphasis to the second formant frequency. A speaker independent HMM based speech recognition system was developed using 30 word vocabulary. The performance of the system was evaluated using neutral, loud, Lombard and angry speech. The recognition rate of mel scale based cepstral coefficients under neutral condition was 85.37% and average recognition rate under the stressed condition was 63.89%. For modified mel scale based cepstral coefficients, the recognition rate for neutral condition was 86.30% and the average recognition rate under stressed condition was 67.04%. For ExpoLog scale based cepstral coefficients, the recognition rate for neutral condition was 83.33% and average recognition rate under stressed condition was 68.66%.

1.2.2.4 Recognition Using Stress Compensation

This approach is used to eliminate the variability present in the speech due to stress during the testing phase. Under stressed condition, feature vectors contain speech information and additional stress information. Chen [11] assumed this stress information as deterministic and additive to the speech information in cepstral domain. The means of cepstral values of stressed speech are shifted from those of neutral speech due to variation in the spectral tilt [11]. Variation in the spectral tilt occurred due to the additional stress information present in the speech signal. An adaptive mean cepstral subtraction technique was proposed, which subtracts the means of estimated cepstral vectors of the stress from those of stressed speech vectors. This compensation technique brings the modified cepstral vectors near to the cepstral vectors of neutral speech. The modified cepstral vectors were considered as input feature for recognition. A speaker independent HMM based isolated word recognizer was developed for 105 isolated words vocabulary. System was trained using neutral speech and during the testing, speech under neutral, fast, loud, Lombard, soft and shout conditions were considered. Average error rate of proposed stress compensation based speech recognition was 9.0%, whereas, the error rate of speech recognizer without compensation was 10.4%. In this study, the stress information

was assumed unchanged at the word level. However, Hansen [12] observed that the effect of stress was uniformly distributed at phoneme level. In this study, the same stress compensation technique was developed for noisy and Lombard speech recognition at the phoneme level. A speaker independent HMM based speech recognizer was developed using 35 word vocabulary. The average recognition rate of noisy Lombard speech was improved from 36.7% to 74.7% using this stress compensation technique. Afify et al. [59] assumed the stress component as additive random bias at the state level in continuous density hidden Markov model framework. In addition, the speech and the stress components are assumed as statistically independent. Maximum likelihood state based additive bias model compensation technique is proposed. In this technique, the independent bias model and the polynomial trend model are used. A speaker independent HMM based speech recognizer was developed using 21 words vocabulary.
Hansen and Clements [60] developed spectral domain stress compensator. In this study, four formant locations and their bandwidth features were considered for stress compensation. Each feature was assumed to have Gaussian distribution and a stress transformation based unique codebook was developed for each phoneme under stressed condition. The stress transformation factor is computed as ratio of sample mean of feature under neutral condition to the sample mean of the same feature under stressed condition. According to stress transformation factor, stressed input speech features transformed into neutral speech features. In this study, a 30 vocabulary speaker independent HMM based isolated word recognizer was developed. The performance of recognizer was evaluated for neutral, angry, Lombard, loud, soft, slow and fast speech. Average recognition rate of this compensation technique based speech recognition was observed to be 57%, whereas, the baseline performance of the recognizer was 30.3%. Here, baseline refers to the system trained with neutral speech and tested with all the stressed speech without considering any compensation technique.

Raja [62] explored four compensation techniques for speaker recognition under stressed condition namely, speaker and stress information based compensation($SSIC$), compensation by removal of stress vectors ($CRSV$), combination of $MFCC$ and sinusoidal amplitude ($CMSA$) features and cepstral mean normalization (CMN). The speaker recognition performance was

Table 1.1: Summary of existing approaches for stressed speech recognition. In the table the abbreviations ER represents error rate and RR represents recognition rate. In case of ER and RR, the values inside the bracket indicate performance without using corresponding approach for speech recognition

Approach	Feature	Database	Performance	Limitation
Multistyle training	Cepstral [51]	(35 words) Neutral, Angry, Lombard, Loud, Soft,Slow and Fast	ER: 9.8% (20.7%)	Performance degrades if test conditions drift from the original trained data
Classification of Stressed Speech	Excitation: Duration,intensity,pitch, cepstral: MFCC, *AC* MFCC [3] Spectral:Slope,formant location,bandwidth [15], subband energies [61]	(35 words) Neutral, Angry, Lombard, Loud, Soft,Slow and Fast	RR: 80.6% (69.5%)	Recognition rate linked to performance of stress classifier
Robust Feature Extraction Technique	Modified Mel scale and ExpoLog scale [10]	(30 words) Loud, angry and Lombard	Mod.Mel scale RR: 67.04% ExpoLog: RR: 68.66% (63.89%)	Applicable for specific stressed condition
Compensation Technique	Formant locations and bandwidths [60] Cepstral mean normalization [11] [12]	(20 words) Neutral, Angry, Lombard, Loud, Soft, Slow and Fast	RR: 57.3% (30.3%)	Effective only for certain stress conditions

evaluated for neutral, angry, question and Lombard conditions. $SSIC$ technique was based on F-ratio test of $MFCC$ coefficient for speaker as well as stress. Higher F-ratio of a coefficient between the speakers per the stress classes represents high variability. Only those $MFCC$ coefficients were considered, which had high F-ratio between the speaker and low F-ratio between stress classes. The average speaker identification rate was improved to 56.34% from a baseline system performance of 53.96%. In $CRSV$ technique, the effect of stress was evaluated by measuring the distance between stressed speech and neutral speech with a predefined threshold. This threshold was decided based on distance between average neutral speech feature vectors and common neutral codebook (training feature vectors from all speakers). Only those feature vectors were considered which crossed the threshold value. The average speaker identification was improved to 54.76% from 53.96%. In $CMSA$ technique the spectral peak was observed to contain speaker specific information that may be suppressed in case of $MFCC$. Therefore a compensation technique using a combination of $MFCC$ and sinusoidal features was developed. The average speaker identification was improved to 54.74% from 53.96%. For a particular stress, the cepstral vectors of stressed speech were normalized with its mean cepstral vector to make it close to the neutral speech cepstral vectors. The average speaker identification was found to be 52.38%. This result showed that the performance does not improve from baseline performance after cepstral mean normalization.

A summary of existing approaches and their techniques is shown in Table 1.1. All mentioned techniques are employed in speaker independent speech recognition except the first work i.e. simulated speech token. In the first case, different utterances of the same speaker were used for training and testing. In most of the cases, either $SUSAS$ database or data collected in their own laboratory are used. Types of stress are also mentioned in the third column. The performance of each approach demonstrates the effectiveness of this techniques under stressed condition. These techniques perform well for certain conditions and for some other conditions, the effectiveness of these techniques are limited. The limitations of these techniques are mentioned in the fourth column. It was observed that the speech recognition performance is related to the classification rate of the stress and it is applicable for limited stressed conditions only.

1.3 Scope of the Present Work

In most of the cases studied in the literature, SUSAS database of 35 words vocabulary is used. In some cases, either a subset of SUSAS database or database collected in their own laboratory environment are used. Since, the databases are different; the performances of these approaches cannot be compared directly. Also, the databases have small vocabulary size and the system is trained and tested with small number of utterances which may not justify the effectiveness of the proposed techniques. A stressed speech database therefore needs to be collected, which have at least moderate size of vocabulary and sufficient utterances for training and testing. The performances of these approaches should be then reevaluated using this database under different stressed conditions.

Literature shows that the $MFCC$ feature performs better than other existing speech features for speech recognition. It is also observed that the recognition rate of $MFCC$ feature is not at par with its own performance under neutral condition, when the recognizer is subjected to the stressed condition. The $MFCC$ feature is derived from the spectral energy of the speech signal, which depends on the spectral features such as spectral tilt [10], formant locations and bandwidths of the spectrum. The variations in these features due to stress influence the spectral energies as well as cepstral coefficients of the speech signal [60]. Hence, a detailed analysis of spectral and cepstral features under stressed condition may provide better insight into the cause of degradation in the recognition performance.

Under stressed condition, the speech signal contains additional stress information. In literature, this stress information is assumed as deterministic and additive at word and at broad phonemes levels. The stress information is also assumed as statistically independent of the speech information. The relation between speech and stress information can also be evaluated for separation of stress from the speech signal.

1.4 Organization of the Thesis

The contents of the thesis are organized as follows

Chapter 2 presents review of analysis of stress using different speech features for stressed speech recognition. Section 2.1 presents first a review of stressed speech databases. Section 2.2 describes the brief review of features and classifier used for speech recognition. This section also discusses the speech recognition performance under stressed condition. The analysis of stress on different features is described in Section 2.3. The motivation of the work is discussed in Section 2.4.

Chapter 3 describes the development and evaluation of stressed speech database. The stress and speech information present in the database are validated by evaluating the stress class and speech information present in the recorded utterances. The stress and speech information are evaluated perceptually as well as by using automatic methods for stress classification and speech recognition, respectively. Section 3.1 describes the recording and development of simulated stressed speech database. Section 3.2 describes the evaluation of stress and speech information of the database using stress classifier and speech recognizer. The performances of existing approaches of stressed speech recognition are described in Section 3.3.

Chapter 4 describes the analysis of stress using spectral tilt. Two methods of computation of spectral tilt such as conventional method and relative formant peak displacement RFD are used to evaluate the stress information. Analysis of stress using conventional method of computation of spectral tilt is studied in Section 4.1. The analysis of stress using the RFD feature is studied in Section 4.2. Section 4.3 presents the evaluation of stress classification using RFD feature. Section 4.4 describes stressed speech recognition using proposed stress classifier.

Chapter 5 presents the analysis of stress using subband energy. The statistical characteristics such as mean, variance and divergence of subband energy are investigated for stressed condition. The difference energy information of subband energy is also investigated for stressed condition. The statistical analyses of the subbands under stressed condition are described in Section 5.1. Techniques to compensate the effect of stress from the subbands are described in Section 5.2. The effectiveness of these compensation techniques are evaluated in Section 5.3. The performance of stress compensation based speech recognition are evaluated in Section 5.4.

Chapter 6 presents analysis of stress using the subspace projection. A subspace projection

based approach is used to separate the speech and stress information from the stressed speech signal. In this approach, the speech and stress information are assumed to be orthogonal. Orthogonality assumption between speech and stress information is verified experimentally. The principle of subspace projection based approach is described in Section 6.1. The orthogonal relation between speech and stress information is verified in Section 6.2. The subspace projection based speech recognition is described in Section 6.3. The proposed techniques for speech recognition application are evaluated in Section 6.4.

Chapter 7 concludes the thesis by summarizing contributions and identifying future scope of the present work.

2

Analysis of Stressed Speech for Speech Recognition-A Review

Contents

Features based on spectral information are widely used for speech recognition. The most successful state-of-the-art features, mel frequency cepstral coefficient ($MFCC$) and linear prediction cepstral coefficients ($LPCC$), mainly capture spectral information of speech signal [4]. The speech recognition system using these features performs well when speech is produced under neutral condition [51]. However, the spectral information of the speech signal varies under stressed conditions which leads to the variations in $MFCC$ and $LPCC$ features from those under the neutral condition. The variation in these features causes degradation in the performance of speech recognition system.

This chapter presents reviews of three broad areas of speech recognition system under stressed condition. First, the review of different stressed speech databases is discussed. The stressed speech database is required to judge the effectiveness of any technique for stressed speech recognition. The second one presents the review of speech recognition system which includes speech features and classifiers used for speech recognition. Then a review of analysis of stress using different speech features is presented. A review of stressed speech database is presented in Section 2.1. Section 2.2 describes the brief review of speech recognition under stressed condition. The analysis of stress on different features is described in Section 2.3. The motivation of this work is discussed in Section 2.4.

2.1 Review of Stressed Speech Databases

Stressed speech database plays an important role for speech processing areas under stressed condition. It provides scope to the researchers to investigate the effect of stress on speech waveform and to use this investigation for the development of robust approach for speech processing applications. Several stressed speech databases are collected for analysis of stress on speech signal. These databases are developed depending on the nature of application, types of stresses, languages and the manner of stress expression [63, 64]. The stressed speech databases are collected for different speech processing purposes such as speech and speaker recognitions, stress classification and stressed speech synthesis [64]. Most of the databases are collected in speaker's own languages such as the English, the German, the Spanish and the Japanese

languages [64]. The most common stresses used for recording are anger, sadness, happiness, fear, disgust, boredom, surprise and joy.

Databases are mainly recorded by three ways to express the stress namely, spontaneous stressed speech, elicit stressed speech and simulated (read) or acted stressed speech [64]. The spontaneous speech is collected, where speakers speak unknowingly in uncontrolled environment such as in real-life situations [65]. Although, this type of speech reflects authentic stresses, the collection of this speech is difficult. For example, the adjustment of microphone from the speaker and the recording environment condition cannot be controlled due to which, the acoustic properties of this speech signal are weak and distorted. Hence, the acoustic analysis on this speech is difficult. An alternative way to collect the stressed speech database is to have a number of speakers elicit stresses while reading prompted texts. The content of text depends on the type of stress to be elicited. From the stress authenticity point of view, these speech contain appropriate stress content. However, the collection of speech with varying study materials may not be an appropriate strategy for content based retrieval application such as speech recognition. The simulation of stress with non-varying lexical content may give useful information to such retrieval applications. Normally, the simulation of stressed speech is done with semantically neutral text speech material collected under controlled environment. In this type of simulation of stress, speakers are aware of vocal expression of stress and language. Although, the simulated stressed speech may not contain authentic stress, the speech can be considered usable for analysis and investigation of stress. Further, the research using such database may give useful information to the content based retrieval application.

2.1.1 Speech Under Simulated and Actual Stress ($SUSAS$) Database

$SUSAS$ database was collected for analysis of stress for speech recognition, which is an application of content based retrieval [66]. The simulated portion of this database was collected to provide scope to the researcher for the development of robust speech recognition system. The database consists of normal, angry, soft, loud, slow, fast, clear, cond50, cond70 and Lombard speech [67]. The cond50 and cond70 speech were recorded where speakers were asked to

engage themselves in tracking tasks under different levels of workload. Lombard speech was obtained by playing 85 dB SPL pink noise to the speaker through headphones, while speaking (i.e., recordings are noisefree). The database was recorded at 8 kHz sampling frequency and 16 bits per sample resolution. Thirty five aircraft communication English words were used for database collection from nine American speakers. Database consists of twelve utterances of each word under neutral condition and two utterances per word for each stressed condition. The complete database comprises of 8820 utterances. The perceptual validity of stress for $SUSAS$ database was evaluated in [16]. Listeners perceived 52.83% of the (in average sense) angry, loud, neutral and slow speech correctly. The $MFCC$ based automatic stress classifier identifies approximately 63.25% stress classes [55].

2.1.2 Berlin Emotional Speech Database

The German database was collected for 10 semantically neutral texts from 10 professional speakers (5 males and 5 females) [68]. The database was recorded in neutral, angry, sad, boredom, joy, fear, and disgust emotions. The complete database comprises 800 utterances. The database was recorded at 48 kHz sampling frequency and 16 bits per sample resolution. This database is mainly used for analysis, classification and synthesis of stressed speech. The perceptual and automatic stress identifications were evaluated for this database. It shows 67.3% and 74.5% stress identifications from listening test and stress classification technique, respectively.

2.1.3 Speech Under Simulated Emotion ($SUSE$)

The $SUSE$ database was collected in two languages namely, the English and Telugu, an Indian language [9]. The database contains two texts, one in English and other in Telugu, spoken from 32 Telugu native speakers. The database was recorded in angry, compassion, happy and neutral conditions. The database was recorded at 16 kHz sampling frequency and 16 bits per sample resolution. The complete database comprises of 600 utterances. This database is mainly used for analysis and classification of stressed speech. The listening test and automatic

stress classification were evaluated for this database. Results showed that listeners are able to identify 80% and 70% stresses in Telugu and English databases, respectively. Similarly, automatic stress classifier is able to identify 55.4% and 56.8% stresses in Telugu and English databases, respectively.

The above mentioned databases can be used for investigation of stress on vocal parameters and these investigations can be used for stress classification and synthesis purposes. However, due to their limited number of utterances for training and testing, the effectiveness of any stress specific analysis cannot be judged for speech recognition application. Hence, a stressed speech database is required for analysis of stress on the speech signal for speech recognition purpose.

2.2 Review of Speech Recognition System

This section presents a brief review of speech recognition system. The basic blocks required for any speech recognition system are the feature extraction, pattern classification and pattern comparison. This section presents a review of each of the basic blocks.

2.2.0.1 Preemphasis

Speech signal contains vocal tract system and excitation source information. Along with these information, the speech signal also contains glottal source and radiation effect information [2]. Glottal source and radiation effect information normally interfere during estimation of vocal tract and excitation source information. Preemphasis suppresses glottal source and radiation effect from the speech signal. Preemphasis filter is high pass filter which emphasizes the high frequency contents. This filter reduces the dynamic range of the frequency content by flattening the spectral tilt which, improves linear modeling of the formant structure [69]. The most widely used preemphasis filter is a fixed first order system whose transfer function $H(z)$ is given by,

$$H(z) = 1 - az^{-1}, \tag{2.1}$$

Typical value of a varies from .9 to 1. In this thesis work, the value of a is kept to be 0.97.

2.2.1 Feature Extraction

In this section, some feature to extract the acoustic properties of the speech signal are reviewed. A detail review of MFCC, LPCC and Delta cepstral features are presented in this section.

2.2.1.1 Mel-frequency Cepstral Coefficients ($MFCC$)

$MFCC$ feature is estimated using critical filters which are based on human perception [4]. These filters are warped on a nonlinear mel-frequency scale which is linearly spaced in low frequency region i.e. below 1000 Hz and it is logarithmically spaced above 1000 Hz [2]. As a result, the modified filterbanks suppress the insignificant spectral variations of the higher frequency bands and capture phonetically important information of speech by giving more importance to the lower frequency bands. For $MFCC$ computation, the speech signal is framed into blocks of samples of speech and the frames of speech signal are processed using window to minimize the effect of discontinuity occurred at the beginning and ending of the frame [2]. In this thesis work, 20 msec frame size and 10 msec frame rate are considered for the speech signal. The discrete Fourier transform (DFT) of windowed frame is taken to convert the time domain signal into frequency domain signal. The frequency domain representation of the signal is termed as spectrum of speech. According to the psychophysical studies, human perception of the frequency content of sounds does not follow linear scale. The subjective pitch is measured for a tone with an actual frequency, f_{Hz}, on a scale called "mel" scale. The relation of actual measured frequency f_{Hz} and subjective pitch (f_{mel}) is given as [2]

$$f_{mel} = 2595 \times log(1 + \frac{f_{Hz}}{700}) \tag{2.2}$$

The subjective nonlinear perception of frequency is used to convert a measured spectrum of given sound into a perceptual spectrum. The perceptual spectrum uses filterbanks which are uniformly spaced on nonlinear warped mel scale. Each filter in filterbanks has a triangular bandpass frequency response. These filters are termed as critical filters. This thesis work considered 24 filterbanks. The modified spectrum, S_k where $k = 1, 2, ...K$, consists of output

powers of K filters. The discrete cosine transform (DCT) of log magnitude of S_k gives the cepstral coefficients of length L as given below

$$c_l = \sum_{k=1}^{K} (logS_k) cos[l(k - \frac{1}{2})\frac{\pi}{K}], \quad l = 1, 2, ...L \tag{2.3}$$

In this thesis work, thirteen cepstral coefficients are computed from each frame of the speech signal. The set of $MFCC$ features of these frames is considered as feature vectors of an utterance.

2.2.1.2 Linear Prediction Cepstral Coefficients ($LPCC$)

Linear prediction coding based features are widely used in speech recognition application [27]. It provides good approximation of vocal tract spectral envelope. In linear prediction (LP) model, a given speech sample $s(n)$ at time n can be approximated as a linear combination of the past p speech samples,

$$s(n) = \sum_{k=1}^{p} a_k s(n - k) \tag{2.4}$$

where, a_k is the prediction coefficient. The LP coefficients are computed from autocorrelation method. $p + 1$ autocorrelation values of each frame are converted to LP coefficients using Levinson-Durbin's method. In this thesis work, ten LP coefficients are computed from each of the frames of the speech signal. The Fourier transform of the log magnitude spectrum gives the cepstral coefficients of the corresponding LP coefficients. The conversion of LP coefficients to $LPCC$ is done using recursion method as given below [2]

$$c_0 = ln\sigma^2 \tag{2.5}$$

$$c_l = a_l + \sum_{k=1}^{l-1} (\frac{k}{l}) c_k a_{l-k}, \qquad 1 \leq l \leq p \tag{2.6}$$

$$c_l = \sum_{k=1}^{l-1} (\frac{k}{l}) c_k a_{l-k}, \qquad l \geq p \tag{2.7}$$

where, σ^2 is the gain of LP coefficients. The cepstral coefficients were observed to be more

robust and reliable feature for speech recognition application as compared to LP coefficients [4], [2].

2.2.1.3 Delta Cepstral Features

The cepstral feature provides good representation of local spectral properties of the speech signal [4], [2]. The first and second derivatives of cepstral feature were observed to perform better for speech recognition as compared to cepstral feature [44]. The delta cepstral features capture dynamic information of the cepstral feature [44]. The delta cepstral feature of l^{th} coefficient at time index n can be defined as

$$\Delta c_l[n] = \sum_{m=-M}^{M} m c_l[n+m] \tag{2.8}$$

where, M is the number of frames over which the delta cepstral feature is computed. In this thesis work, M is kept to be 2. Similarly, delta-delta cepstral feature can be defined as delta operation on delta cepstral features. In case of $MFCC$ feature of 13 coefficients, the combination of static and dynamic cepstral features become $MFCC + \Delta + \Delta\Delta$ feature of 39 coefficients. This feature shows good recognition accuracy in case of noisy speech recognition [44].

2.2.2 Classifiers for Speech Recognition

Various modeling techniques were explored in the literature to model features for speech recognition. Mostly used classifiers include VQ [70], HMM [30] and ANN [32]. Among these, HMM is considered as most successful classifier for speech recognition [30]. In this thesis work, HMM based classifier is used to evaluate the effectiveness of these features.

2.2.2.1 Hidden Markov Model (HMM)

Hidden Markov model approach is widely used as a statistical method to characterize the spectral properties of a pattern. This technique is applicable for those signals and processes which are stochastic in nature, that is, their statistical characteristics vary with time. HMM is a doubly stochastic process with an underlying stochastic process, which is hidden. It can only

be observed through another set of stochastic process, that generates the sequence of observed symbols [71]. HMM can be characterized by its number of states in the model, number of distinct observation symbols per state, the state transition probability distribution, the observation symbol probability distribution, and the initial state distribution. HMM deals with the observations, which are characterized either by discrete or continuous signals. The discrete signal is modeled using discrete probability density within each state. The continuous observation can also be characterized using discrete density model by converting the continuous signal into discrete observation sequence by using any quantization technique. The continuous signal can be modeled using continuous observation density. The continuous observation density can also be modeled using probability density function reliably. The observation symbol probability, $(\mathbf{b}_j(\mathbf{o}))$, can be written as

$$\mathbf{b}_j(\mathbf{o}) = \sum_{k=1}^{M} c_{jk}\mathcal{N}(\mathbf{o}, \boldsymbol{\mu}_{jk}, \mathbf{C}_{jk}), \qquad 1 \leq j \leq N \tag{2.9}$$

where, $\mathbf{o}$ is the observation vector, c_{jk} is the mixture coefficients of k^{th} mixture in the j^{th} state. In speech recognition, $\mathcal{N}$ is assumed as Gaussian with mean vector $\boldsymbol{\mu}_{jk}$ and covariance matrix $\mathbf{C}_{jk}$ for the k^{th} mixture component in state j.

The HMM can be classified by the structure of the transition matrix, $\mathbf{T}$, of Markov chain. One type of HMM is based on ergodic model that has property to reach every state from any other state in a finite but aperiodic number of steps. The other type of HMM is left to right or Bakis model. It has property that as time increases, the state index increases, that is, the system state proceeds from left to right. This type of model is useful for those signals whose signal properties change over time in a successive manner. In speech recognition, left to right model is widely used. In this model, the state transition probability has property that no transition is allowed to those states whose indices are lower than that of the current state.

$$t_{ij} = 0, \qquad j < i \tag{2.10}$$

The initial transition probability has property that state sequence must begin in state 1, i.e.,

$$\pi_i = \begin{cases} 0 & i \neq 1 \\ 1 & i = 1 \end{cases} \tag{2.11}$$

2.2.3 Speech Recognition System

In this section, isolated word recognizer is demonstrated using HMM classifier. The objective of speech recognizer is to identify the spoken message present in an utterance. The model is characterized by N number of states and M number of distinct observation symbols, $v_1, v_2, \ldots, v_M$ per state. The initial state distribution $\boldsymbol{\pi} = \{\pi_i\}$, in which π_i, $1 \leq i \leq N$, is the probability of occupying state i, at time $t = 1$. The transition matrix, $\mathbf{T} = [t_{ij}]$ is the probability of occupying state j at time $t + 1$ from state i at time t. $\mathbf{B} = [b_{jk}]$ is the probability of symbol v_k in the state j. Thus, the model of a word is defined by the three parameters, $\lambda = (\boldsymbol{\pi}, \mathbf{T}, \mathbf{B})$. The model parameter is obtained by optimizing the likelihood of observation vectors of the training set for the w^{th} word. A set of word models, $\{\lambda_1, \lambda_2, \ldots, \lambda_W\}$ can be developed for W words vocabulary using this procedure.

The utterance, $\mathbf{o}$ of a word consists of a sequence of T observations, $\mathbf{o} = o_1, o_2, \ldots, o_T$ obtained via feature analysis of the speech. The recognition task determines among W word models, which model is the most likely to have produced $\mathbf{o}$ as given below

$$\hat{w} = \underset{1 \leq w \leq W}{\arg\max}[P(\mathbf{o}|\lambda_w)] \tag{2.12}$$

The probability computation is done using Viterbi algorithm, where maximum likelihood path is considered. The Viterbi algorithm reduces the computational complexity.

In the present work, $SUSAS$ database is used to demonstrate the isolated word recognition. Table 2.1 shows speech recognition results for $SUSAS$ database [66]. In this database, twelve utterances of each word are considered for training and two utterances per word are considered for testing. The word model is developed using continuous density HMM. Ten states are considered to model a word and each state has two mixture components. Speech recognition system is developed using $MFCC$ and $LPCC$ features and the recognition performances of

Table 2.1: Performance of speech recognition of *SUSAS* database for different features (in%)

Speech features	Recognition rate
$MFCC$	74.92
$LPCC$	73.02
$MFCC$+Preemphasis	77.94
$LPCC$+Preemphasis	75.71
$MFCC + \Delta + \Delta\Delta$	94.07
$LPCC + \Delta + \Delta\Delta$	93.15

these features are given in Table 2.1. The recognition performances of $MFCC$ and $LPCC$ features, without using preemphasis, are observed to be 74.92% and 73.02%, respectively. This result infers that in comparison to $LPCC$ feature, the $MFCC$ feature captures speech information more accurately. The performances of $MFCC$ and $LPCC$ features are improved to 77.94% and 75.71%, respectively, when preemphasis filter is applied to the speech signal. The performances of $MFCC$ and $LPCC$ features, along with their derivatives are further improved to 94.07% and 93.15%, respectively. These results show that $MFCC$ feature performs better than $LPCC$ feature when speech is neutral. In order to verify the effectiveness of these features for recognition of speech under stressed condition, the performances of these features are evaluated under stressed condition and their performances are shown in Table 2.2. The performances

Table 2.2: Performance of speech recognition of *SUSAS* database for different stressed conditions (in%)

Speech features	Neutral	Angry	Slow	Loud	Lombard
$MFCC$	77.94	39.72	65.24	51.66	51.77
$LPCC$	75.71	33.39	59.37	44.60	34.55

of $MFCC$ feature under angry, slow, loud and Lombard conditions are observed to be 39.72%, 65.24%, 51.66% and 51.77%, respectively. Similar observation can be made for $LPCC$ feature. These results infer that the performances of these features under stressed conditions are not at par with their own performances under neutral condition. These results also supported that the characteristics of the speech signal vary under stressed condition, due to which, the feature

of stressed speech are different from that of neutral speech. Hence, a detailed analysis of stress on different characteristics of speech may give useful insight of stress information, which may be useful to increase the robustness of speech recognition system.

2.3 Analysis of Stress Using Different Speech Features

This section gives a review of stress analysis on three speech features:

(i) Temporal feature

(ii) Spectral feature

(iii) Cepstral feature

2.3.1 Temporal Features

The typical temporal parameters used for stress analysis are duration, intensity, fundamental frequency (F_0), and glottal parameters [67]. Duration of an utterance or sound units such as phonemes, syllables or words, can be defined as the time taken to complete utterance or sound units. Fairbank and Hoaglin [72] analyzed duration of speech for anger, fear, indifference, contempt and grief. According to this study, the duration of utterance depends on speaking rate (number of words per minute), length of pause and ratio of pause duration to the total phonation time [72], [73]. The duration of utterances spoken under angry, fear, indifference conditions were observed to be shorter whereas, the duration of utterances under contempt and grief conditions were observed to be longer. Williams and Stevens [74] analyzed duration of the utterance for angry, fear and sorrow conditions. According to this study, the duration of an utterance depends on the respiration rate. Increased respiration rate may lead to shorter duration of utterance. The duration of utterances spoken under fear condition were observed to be longer than those under angry condition. Hansen and Patil [75] analyzed stress on duration of words and subwords such as vowels, consonants, semivowels and diphthongs. The duration of consonants, semivowels and diphthongs were observed to be approximately constant for

slow and loud speech, whereas vowel duration of soft speech decreased and it was increased significantly for angry and loud speech.

The intensity of speech is the root mean square (RMS) energy of the speech signal. Williams and Stevens [74] observed higher intensity for angry speech compared to the neutral speech. Heulf et al. [76] and Iida et al. [77] observed that the mean intensity of speech under angry, fear and joy was higher than those under disgust, sad and neutral conditions. Hansen and Patil [74] analyzed intensity of words, consonants and vowels under different stressed conditions. The overall intensity were observed to remain constant for vowels under slow and Lombard conditions, whereas, intensity of consonants increase for soft and angry conditions.

F_0 can be defined as average rate of vibration of vocal cords. Williams and Stevens [74] analyzed F_{0med} (median of F_0) and F_{0range} of utterance for angry, fear and sorrow conditions. According to this study, the F_0 of an utterance depends on subglottal pressure, which depends on the respiration rate. The increased subglottal pressure may lead to increase F_0 of utterance. It was observed that the F_{0med} and F_{0range} of utterance under angry condition increased whereas, F_{0med} and F_{0range} of utterance under sorrow condition reduced. In another study, Murry and Aunott [78] analyzed F_{0avg} (average of F_0) and F_{0range} of speech under angry, sad, happy, fear and disgust conditions. In this study, F_{0avg} was observed as highest for angry speech and lowest for disgust speech. F_{0range} was observed to be shorter for sad speech.

The glottal parameters also contain stress information [79, 80]. These parameters are opening slope, closing slope, opening duration, top duration, closing duration, and closed duration [79, 81]. These parameters control glottal air flow. The slopes of glottal opening and closing provide control over the amount of acoustic energy produced and the durations of glottal closure decide the length of pitch period. Cummings and Clements [79] analyzed ratios of closing to opening slopes and closing to opening durations for normal, angry, loud, soft, slow, fast, clear, question and Lombard speech. Study showed that most of the stress information are present in the opening slope, closing slope and closed duration. For example, the ratio of closing to opening slope of speech significantly varied for angry, fast, loud, question, and soft stresses. Similarly, angry, loud, and soft speech showed different closing to opening duration

ratios.

2.3.2 Spectral Features

Typical spectral parameters used for stress analysis are formant locations and bandwidths, strength of formant location, spectral slope and spectral energy [2]. Formants are the quantitative representation of vocal tract, which contain the phonetic information of sounds. The stress modifies the shape and the cross-sectional areas of vocal tract which may be reflected from formant locations and bandwidths [3], [64]. Yildirim et al. [82] observed that F_2 is confused with F_1 for angry and happy speech and F_1 interferes with F_0 under these conditions. Hansen and Patil [75] observed significant shift in F_1 for slow, angry, loud and clear speech. F_2 increases in most of the conditions and slight change was observed for F_3 and F_4 across all stressed conditions. Formant bandwidths for F_1 and F_2 showed large variation and some variation was observed for F_3 and F_4. The strength of formant location named as formant peak plays an important role in vowel perception. Seshadri and Yegnanarayana [13] observed abrupt closing phase of glottal cycle for angry, loud and Lombard conditions which produce variation in perception of loudness. The variation in the loudness may lead to the variation in formant peaks.

Glottal parameters influence the spectral energy of speech [79, 81]. Hansen [81] analyzed glottal parameters under slow, fast, soft, loud, angry, clear, question and Lombard conditions. Under certain stressed conditions, such as angry, loud and Lombard, glottal pulses have sharp rise times and sharp corners. As a result, the high frequency content is relatively increased from that of lower frequency in the spectrum. The relative variation in frequency content produces migration of spectral energy [1]. The migration of spectral energy influences the spectral tilt and the energy of subbands, which are the output of filterbanks placed in the spectrum. Tartter et al. [17] considered spectral tilt as a measure to investigate the distribution of spectral energy of Lombard speech from lower formants to higher formants. In this study, spectral tilt was measured as a slope of a regression line fitted to the spectrum. Klatt and Klatt [83] observed that the change in spectral tilt gives information about variation in speaker. Hansen [81]

analyzed spectral tilt for slow, fast, soft, loud, angry, and Lombard conditions and observed that the spectral tilt varies significantly for angry, loud and Lombard conditions compared to the soft and the slow conditions. Lu and Cook [84] observed that the intelligibility and the loudness of the speech are increased due to increase in flatness of spectral slope for Lombard speech.

Stanton et al. [85] analyzed stress information in energy of uniformly placed filterbanks for neutral, loud and Lombard speech. For vowels, the energy in the frequency range of 500-4000 Hz increased under loud and Lombard conditions and for frequency ranges 0-500 Hz and 4-8 kHz, the energy under these conditions decreased. This study showed that the additional spectral energy moves from low to midbands, which inferred that the migration of spectral energy takes place [85]. For fricative, the shift in energy takes place at frequencies 4000-8000 Hz. Hansen and Womack [10] analyzed the migration of spectral energy in uniformly placed subbands for angry, loud and Lombard speech. The maximum recognition performance was observed near 2^{nd} formant region rather than 1^{st} formant region for these speech. Sarikaya and Gowdy [61] analyzed the stress on individual subbands for slow, fast, soft, loud, angry, question, Cond50/70, neutral, clear and Lombard conditions. It was observed that the temporal deviation of subbands across frames contain stress information.

Ramamohan and Dandapat [9] analyzed stress on amplitude, frequency and phase of sinusoidal based features for neutral, angry, happy, compassion speech. Study showed that the mean amplitudes of speech increased under happy and compassion and decreased for angry speech. On the other hand, frequency of speech increased under angry and happy emotions and decreased for compassion speech. Phase feature did not show much variation across the emotions.

2.3.3 Cepstral Features

The cepstral parameters include $MFCC$, $LPCC$ and log-frequency power coefficients ($LFPC$) features. $MFCC$ feature represents the spectral variation of acoustic speech signal [64]. $MFCC$ feature captures the difference in energy between spectral bands, which is introduced

as a result of variation in vocal tract structure due to stress. Womack and Hansen [3] analyzed stress on $MFCC$, Δ and $\Delta\Delta$ of $MFCC$ and autocorrelation of $MFCC$ (AC_i) features. These features were investigated for angry, neutral, fast, question, slow, clear, Lombard, soft, apache and loud speech. Δ and $\Delta\Delta$ of $MFCC$ provide measure of 'velocity' and 'acceleration', respectively, of $MFCC$ feature. Δ and $\Delta\Delta$ of $MFCC$ features were found to be robust to the stress. AC_i provide measure of correlation and relative change in spectral energy over an extended window frame. The mean and standard deviation of AC_i were evaluated. It was shown that mean and standard deviation of AC_i feature contain stress information. Nwe et al. [86] observed that $LFPC$ feature includes pitch information therefore, it contains better stress information compared to $MFCC$ feature. The $LFPC$ feature was derived by filtering the spectrum with bandpass filter having center frequencies corresponding to the critical bands of the human ear.

Chen [11] analyzed mean cepstral shift for each cepstral coefficient for soft, shout and average of fast, loud and Lombard speech. The mean cepstral shift was the cepstral mean of given stressed speech minus the cepstral mean of neutral speech. Study showed that average shift of the mean values are increased for soft speech and they are decreased for shout and average of fast, loud and Lombard speech. The shift in mean values of cepstral coefficients was observed due to the variation in slope of the spectrum (spectral tilt). This study showed that the migration of spectral energy also effects the cepstral coefficients of the speech signal.

2.4 Motivation for Present Work

The review of stressed speech databases suggests that a stressed speech database is required to investigate the stress specific analysis on speech waveform. The existing techniques of stressed speech recognition are mostly evaluated on $SUSAS$ database. This database has 12 utterances for each word under neutral condition and 2 utterances per word for different stressed conditions. The database is recorded from 9 speakers. Therefore, the recognition accuracy of a word of this database can be given as

$$Acc_{recog} = \frac{1}{(2 \times 9)} \times 100 \tag{2.13}$$

which is approximately 5.56%. Thus, any single correct recognition of the utterance increases the recognition performance by 5.56%. Therefore, database should have more number of utterances for training and testing. Hence, there is a need to develop a new database with sufficient amount of data for training as well as for testing.

It has been observed from the literature that the migration of spectral energy takes place when speech is produced under stressed condition. The migration of spectral energy effects the spectral tilt, the subband energy and the cepstral coefficients. It was also seen that the stress produces variation in the vocal loudness which effects the peaks at formant locations. A detailed investigation of spectral and cepstral features under stressed condition is required for development of robust speech recognition.

The spectral tilt gives information about variation in the speaker such as stress. The analysis of spectral tilt for stressed speech may give useful information about stress. Conventionally, spectral tilt was measured as a slope of a regression line fitted to the spectrum. This technique may capture the gross spectral energy information of the speech signal. Koolagudi and Krothapalli [15] defined spectral tilt as a positive angle between the line joining the 1^{st} formant peak and the 2^{nd} formant peak with the abscissa. All these studies have shown that spectral tilt contains not only gross energy information but also the formant peaks information. The conventional technique for the computation of spectral tilt may not capture the variation in spectral tilt introduced due to variation in formant peaks. Hence, a new method may be required to quantify the variation in the spectral tilt introduced only due to variation in formant peaks.

The spectral energy of the speech signal contains phonetically relevant information of the speech. The migration of spectral energy effects the subband energy. It has been shown in the literature that the individual subbands are effected due to stress and the effect is not uniform across the subbands. It would be interesting to analyze each subband independently when speech is produced under stressed condition.

The migration of spectral energy also effects the coefficient of cepstral features which causes degradation occurred in the speech recognition performance. Under the stressed condition, the

speech signal contains additional stress information. Various stress compensation techniques are proposed in the literature to remove the stress information from the speech signal [11], [12], [59]. These techniques assumed the stress information as deterministic and additive at word and broad phonemes levels [11], [12]. Also, the stress information was assumed as statistically independent of the speech information [59]. It would be interesting to explore the relation between speech and stress information to separate the stress information from the speech signal.

In general, the compensation technique requires prior knowledge of the stress class. Several stress classification techniques were explored to identify the stress classes [8] [9]. The analysis of stress on spectral features showed that spectral features may have stress discriminating ability. Hence, spectral features can be explored for stress classification. The stress classification technique can be used for speech recognition. Some literatures showed that the speech recognition performance depends on the classification rate of the stress classifier [3]. Hence, there is a need to explore techniques for stressed speech recognition which does not depend on the classification rate of the stress classifier and it can be useful for all the stressed conditions. The proposed investigations of this thesis are as follows:

The first one is the development of a stressed speech database. The speech and stress information are evaluated using listening test and automatic methods, to validate the speech and stress information of this database. The performances of different speech features are evaluated using this database and the relative variation in the performances of speech features for stressed conditions with respect to neutral condition is evaluated.

The second one is the investigation of the stress information in the spectrum, where spectral tilt is considered as a feature. The stress discriminating capability is evaluated by conventional spectral tilt method. This study also evaluates variation between formant peaks by defining local spectral tilts as a feature. The stress information on local spectral tilts are evaluated by developing a stress classifier, where local spectral tilt is considered as feature.

The stress information on spectral energy, subband energy of speech is considered as third investigation. The statistical characteristics of subband energy are investigated for different stressed conditions. This study also evaluates the variation of energy across the subbands by

defining difference energy as a feature. All these analyses are used for development of a suitable stress compensation technique. The effectiveness of stress compensation technique is evaluated using speech recognition.

In general, stress compensation technique based speech recognition system requires stress classifier to give explicit knowledge of stress class. The performance of speech recognition system limits due to the classification rate of stress classifier. Hence, as a fourth investigation, a stress compensation technique, which does not require the explicit knowledge of the stress class. A subspace projection based technique, is considered to separate the speech and the stress information from the stressed speech signal. In this technique, an orthogonal relation is assumed between the speech and the stress components. This study evaluates orthogonal assumption using the cepstral feature. The stress compensation technique is developed based on subspace projection based approach. Finally, the performances of all the proposed techniques are evaluated and compared for speech recognition.

3

Simulated Stressed Speech Database: Development and Evaluation

Contents

In order to develop robust speech recognition system, investigation of stress information on different speech features is required. Different approaches are required to model the stress specific information in such a way that it can increase the robustness of the recognizer under stressed condition. A stressed speech database is required that provides the scope for analysis of stress. Several databases have been collected for analysis of stress. The most recent database, *SUSAS* database, is collected to aid the development of robust speech recognition system [66]. The analysis and modeling of speech under stressed condition using *SUSAS* database may not be effective due to its limited number of utterances for training and testing. Hence, there is need for development of a stressed speech database.

The ultimate goal of developing a stressed speech database is to analyze the stress information. This stress specific information can be used for speech recognition application. In this chapter, the development and evaluation of stressed speech database are presented. Since the database has to be used for analysis of stress for speech recognition, which is an application of content based retrieval, therefore, a simulated stressed speech database is required. Normally, this type of database is collected from professional speakers, where it is assumed that the speaker can vocally produce the stress. These speakers produce full blown stress that makes it easier to identify the stress from speech. German and Danish databases are collected using professional speakers [68, 87]. The stress classification system developed using these databases gives good classification accuracy. However, the same classification system fails miserably when the speech is produced by nonprofessional speakers [88]. Thus, any speech processing application developed for these databases may not be effective in a nonprofessional scenario. *FAU* Aibo Emotion Corpus database are collected from nonprofessional speakers [88, 89]. Here, speakers are children whose ages are below 13 years. From speech technology usability point of view, adult non-professional speakers are more important. Hence, there is a need to collect the stressed speech from non-professional adult speakers. This chapter evaluates the stress information present in the acoustic features such as duration, intensity, fundamental frequency and formant frequencies of the speech. This chapter also evaluates how well the stress in the stressed speech is identified by the listeners and the automatic stress classification method using

this database. Both these studies will validate the stress information of the present database.

Section 3.1 describes the recording and development of simulated stressed speech database. Section 3.2 describes the evaluation of stress and speech information of the database using stress classifier and speech recognizer. The performances of existing approaches of stressed speech recognition are described in Section 3.3. Section 3.4 describes the summary of the chapter.

3.1 Development of Speech under Simulated Stress Condition ($SUSSC$) Database

The first step to develop a stressed speech database is to select the types of stresses and the manner of stress expression. In the present work, the database is collected for two psychologically induced stresses, one external surrounding induced stress and the neutral speech. The psychologically induced stresses are angry and sad [90]. The external surrounding induced stress is Lombard speech [12]. The neutral speech is considered as referent. Among the three manners to express the stresses as mentioned in the Chapter 2.1, the simulated speech is chosen for database that would allow controlled analysis of stress on utterance.

3.1.0.1 Selection of Text Corpus

Semantically neutral text is chosen to collect the database, so that the text by itself does not evoke any stress [91]. The database consists of isolated words and continuous texts of Hindi, an Indian language. The selection of these texts is based upon the criteria that the database should contain most of the Hindi language phonemes. In order to study the effect of stress on the continuous speech, some isolated keywords, short phrases, long phrases and passages are formed. There are 33 isolated words, 28 short phrases, 29 long phrases and 3 passages present in the database. The complete text material contains 92 utterances (approximately 730 words) per speaker per stressed condition. The speaker takes approximately 12 to 15 mins to complete this text material for a stress class. The vocabulary of this database is 165 words. One hundred nineteen (119) words are segmented manually from phrases and passages to understand the effect of stress in continuous speech. These words are shown in Table 3.1.

Table 3.1: List of keywords for stressed speech analysis

angoothi	saamaajik	jaankari	ghatana	majduri
darwaaja	bataya	banaaya	videshi	pareshaan
shakaahaari	parivartan	jaayega	lagaataar	saathiyo
puraana	dikhaayi	bhaavana	tumhaari	mariyaada
namaste	kahaani	kitane	samaachar	nivedan
sahaayata	maataaji	parampara	janvari	diya
gaya	nahi	karne	tumhe	mere
barish	bahar	uski	khane	karna
sunaya	jaise	kiya	rakhkar	khana
liye	rahi	raha	karke	logo
liya	mahaaraaja	dakghar	tiket	nikaalkar
laga	garib	mile	rahe	usne
aapse	baje	ghati	samjho	bate
suno	sunata	pani	pahnana	sabhi
mera	bana	rakhunga	palan	maine
kaha	batana	aadat	karun	aapne
aayegi	kare	karenge	sakte	warna
milta	mila	aaya	pehli	samjha
milna	jaane	lane	ghatne	galat
dene	laya	aane	unki	wala
ghate	hote	taki	aapko	dawra
paisa	bheja	iske	lene	gaye
unhe	beghar	diye	raja	milegi
suna	banega			

The phonemes covered by this word set are given below.

(i) Vowels: /a/, /e/, /i/, /u/, /o/

(ii) Consonants:

Velar	/k/	/kh/	/g/	/gh/	
Palatal	/ch/		/j/	/jh/	
Alvelar	/T/		/D/		
Dental	/t/	/th/	/d/		/n/
Bilabial	/p/		/b/	/bh/	/m/

(iii) Semivowels: /y/, /r/, /l/, /v/

(iv) Fricatives: /s/, /sh/, /h/

The database contains single word utterances where words are either isolated or segmented from the phrases and passages.

3.1.0.2 Recording Setup

Speech under Simulated Stress Condition ($SUSSC$) Database is collected at semi-anechoic studio at Electro Medical and Speech Technology Laboratory, Department of Electronics and Electrical Engineering, IIT Guwahati. During the recording, the speakers are asked to read the text from printed papers. Speech is recorded from the table top microphone and the microphone is kept at approximately 10-15 cm distance from the speakers. The gain of the recording phone is kept constant throughout recording to prevent any intensity variation under different stressed conditions. Recording is done with a high quality condenser microphone and the whole session is recorded at 16 kHz sampling rate and 16 bits/sample resolution. The complete database is recorded in two separate sessions in order to prevent it from influencing other speaking styles. The database is recorded by 10 male and 8 female non-professional speakers. The ages of the speakers range from 25 to 40 years. Before recording, a brief introduction of the recording and the text are given to the speakers and the speakers are asked to use their everyday way of expression rather than stage performance [92]. Initially, the speakers are asked to speak all

the utterances in normal mood. Then, the speakers are told to think of a contextual situation, where respective stressed conditions like angry and sad speech are elicited. All the utterances are recorded with one stress at a time. Lombard speech is recorded by simulating acoustic noisy condition. In the beginning of the recording of Lombard speech, the speaker is asked to adjust the speech to a comfortable level (intensity as same as neutral level of speech). During recording, a white noise is played through the headphone of the speaker [12]. The speech signal is stored using wavesurfer software. Each file is coded by its speaker's number, language, stressed condition, word taken from keyword/ short phrase/ long phrase and vocabulary number. For instance, the "/*anguthi*/" word taken from isolated word under angry condition from the first speaker can be coded as "01-AHANKW01.wav".

3.2 Evaluation of Stressed Speech Database

In the present section, the presence of stress in *SUSSC* database is evaluated using acoustic features. This section also evaluates how well the listeners and the automatic method are able to classify utterances in this database according to stressed condition. This study will give an understanding about the level of stress present in the utterance. This section also investigates the ability of the listeners and the automatic method to recognize the content of speech when speech is produced under the stressed condition. Both these studies will help validate the stress and speech information of *SUSSC* database.

3.2.1 Analysis of Acoustic Characteristics of Stressed Speech

In this section, the presence of stress is evaluated using acoustic features [9, 93]. Figure 3.1 shows a recorded speech /*anguthi*/ from a speaker in neutral and angry conditions. This figure compares the variations introduced in the features due to stress. These features are fundamental frequency, duration, and intensity of speech under neutral and angry conditions. From the figure, it is observed that the duration of the speech reduces in angry condition. The average intensity and the fundamental frequency (F_0) of the speech are higher for angry condition.

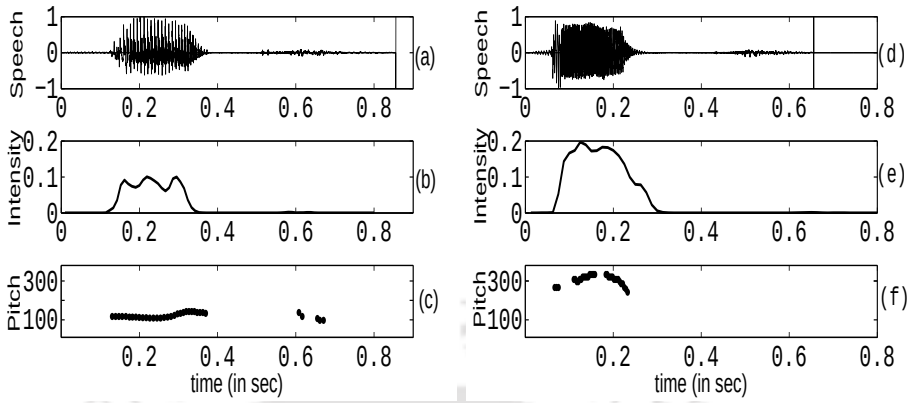

Figure 3.1: Speech signal */anguthi/* recored from one speaker in two different stressed condition (a), (b), (c), neutral condition and (d), (e), (f) angry condition

To get the general trend of variations of these features under different stressed conditions, approximately 1000 utterances per stressed condition are considered. Duration of an utterance can be defined as time taken to speak that utterance. Therefore, approximately 1000 durations per stressed condition are computed. In the present study, duration of an utterance (in msec) is measured as ratio of the number of samples in that utterance to the sampling frequency at which the speech is recorded. The mean and standard (std.) deviation values of duration under different stressed conditions are shown in Figure 3.2. Duration of the utterance depends on the speaking rate [74]. Under angry condition, the speaking rate is higher and as a result, the mean of duration of speech is reduced. On the other hand, under sad and Lombard conditions, the speaking rate is lower, due to which, the mean values of duration of these speech are lower than that of neutral speech. The std, deviation values of duration are also observed to vary with stressed condition. For instance, the std. deviation value reduces for angry speech and increased for sad and Lombard speech.

The fundamental frequency, F_0, of the speech is computed by considering the frame size of 20 msec and frame rate of 5 msec. The F_0 of each frame is estimated by zero-frequency filtering method [94]. Approximately, 50000 frames per stressed condition are considered for evaluation of F_0 values under stressed condition. Mean and std. deviation values of the F_0 values of the speech signal under different stressed conditions are shown in Figure 3.3. The F_0

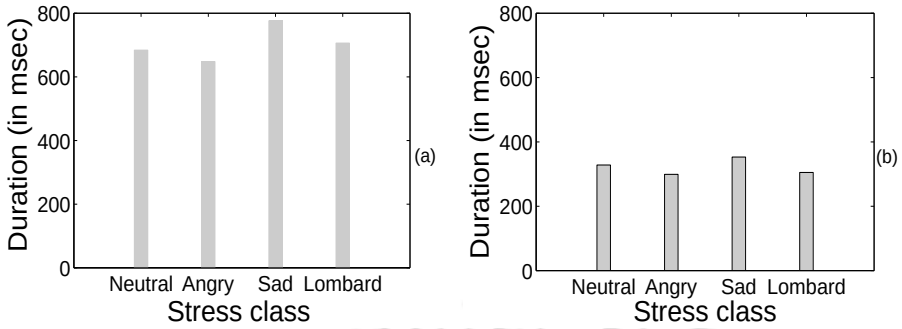

Figure 3.2: Duration of speech under different stressed conditions (a) Mean of duration and (b) std. deviations of duration

value of speech depends on the rate of vibration of vocal folds [74]. Under angry and Lombard conditions, the rate of vibration of vocal folds are higher and as a result, the means of F_0 values of speech are observed to be comparatively higher than that of neutral speech. For sad speech, the rate of vibration of vocal folds is lower and due to which, the mean of F_0 are lower than that of neutral speech. It can also be observed that the std. deviation values of F_0 for angry and Lombard speech are higher and std. deviation value of F_0 for sad speech is lower than that for neutral speech.

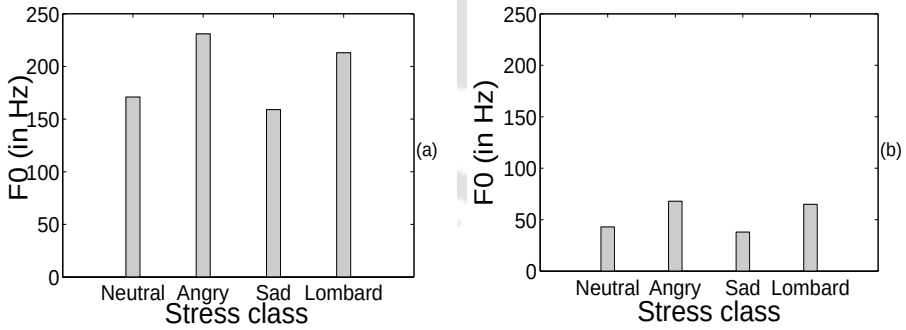

Figure 3.3: F_0 of speech under different stressed conditions (a) Mean of F_0 and (b) std. deviation of F_0

The root mean square (RMS) energy is measured in the frame size of 20 msec and frame rate of 10 msec. Approximately, 25000 frames per stressed condition are considered for evaluation of RMS energy values under different stressed conditions. The mean and std. deviation values of RMS energy of speech under different stressed conditions are shown in Figure 3.4. The

RMS energy value of the speech depends on the rate of vibration of vocal folds and the vocal intensity of speech [74]. Figure shows that the mean values of speech under angry and Lombard conditions are higher than that under the neutral condition. The mean of RMS energy value for sad speech is lesser than that for neutral speech. It can also be observed that the std. deviations of RMS energy under angry and Lombard conditions are higher and std. deviations of RMS energy under sad speech is lower than that for neutral speech.

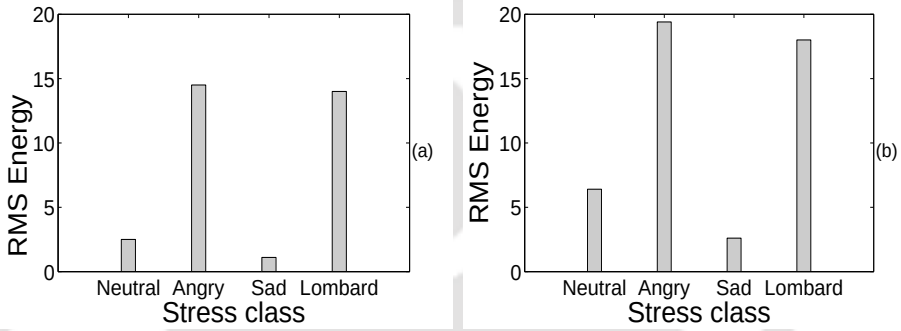

Figure 3.4: RMS energy of speech under different stressed conditions (a) Mean of energy and (b) std. deviation of energy

First four formant frequencies namely, F_1, F_2, F_3, and F_4 of speech are computed from the LP spectrum. LP spectrum of each frame is computed using frame size of 20 msec and frame rate of 10 msec. For vowel region, formant frequencies are more reliable than consonants [2]. Therefore, in the present study, formant frequencies are computed from the vowel regions of the speech signal. These vowels are /a/, /e/, /i/, /u/. Approximately 5000 frames of each vowel under stressed condition are considered to compute the formant frequencies. The mean values of the formant frequencies of vowels under different stressed conditions are shown in Figure 3.5. It shows that the formant frequencies under the stressed condition vary from that of neutral condition. The variations of these frequencies are not uniform for all vowels in stressed condition. Under angry and Lombard conditions, F_1 values of all vowels are higher. F_2 and F_3 values are higher for vowels /a/ and /e/ and these values are lower for vowels /i/ and /u/. F_4 values reduce for all vowels under these conditions. For sad speech, F_1 and F_2 values are lower for all vowels. F_3 value reduces for vowel /a/ and it is higher for /e/, /i/ and /u/. F_4

value is higher for all vowels under sad condition. In general, the first two formant frequencies are higher for angry and Lombard conditions and the other two formant frequencies are lower for these stressed conditions. For sad speech, the first two formant frequencies are lower and other two formant frequencies are higher than those of neutral speech. Further, the variations of mean values of F_1 and F_2 are higher across the stressed conditions compared to those of F_3 and F_4. The trend of variation of these features under the stressed conditions using this

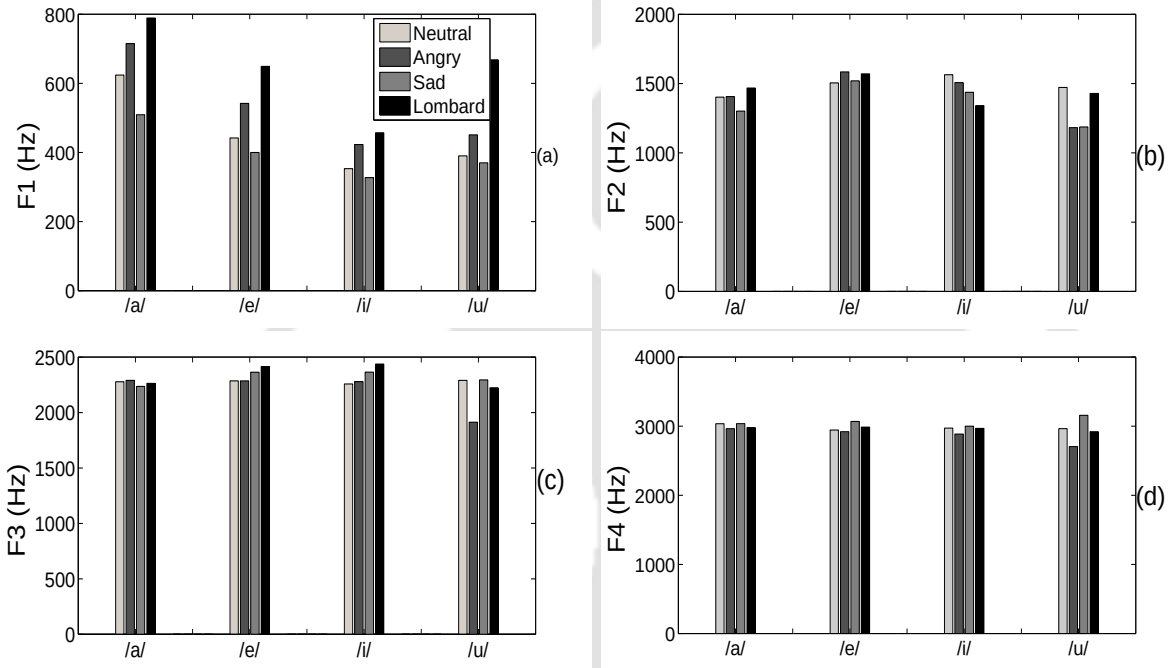

Figure 3.5: Means of formant frequencies of speech under different stressed conditions (a) F_1, (b) F_2, (c) F_3, and (d) F_4

database is observed to be approximately similar as described in Section 2.3. This study infers that this database contains approximately same stress information as other stressed speech databases contained. Hence, the analysis of stress using this database will give approximately the same stress information.

3.2.2 Evaluation of Stress Information

This section presents the evaluation of stress level in the utterance by listening test and automatic stress classification method. In listening test, listeners have to identify the type of

stress present in the utterance. The level of stress present in the utterance of this database is evaluated by the listeners by a method similar to that suggested by Slyh and Bolia [16] and Ramamohan and Dandapat [9]. Out of 119 words, first 33 words from eighteen speaker's speech under four stressed conditions i.e. neutral, angry, sad and Lombard are considered. Each set contains 126 words (18 speakers $\times$ 7 keywords from different stressed conditions). All the files are coded and arranged in random order so that listeners cannot get prior knowledge of stress.

3.2.2.1 Human Perceptual Evaluation of Stress

Seven males and eight females with ages ranging from 25 to 40 years are considered as listeners in this experiment. In this experiment, listeners are chosen to be those persons who are fluent in Hindi language. Fifteen listeners are asked to identify the stress present in the utterance. The listening test were performed by those listeners who were not considered in the stressed speech database collection. Listeners listened all the files and judged via four alternative, forced choice response whether each file is neutral, angry, sad and Lombard. Listener can listen each file until their decisions are finalized. A response of the listener is called correct if the listener's classification matches with its stress label in the *SUSSC* database. The average correct stress classification for the isolated keywords and the segmented keywords are given in the form of confusion matrix. The correct stress classification of listeners for isolated words and segmented words are tabulated in Table 3.2. From this table, it can be observed that the trend of stress classification of isolated words and segmented words are approximately same. This observation infers that the listeners are equally able to identify the stress class in the isolated and segmented words. Therefore, the analysis of stress on segmented words may give similar stress information as that of isolated words. The listeners are able to perceive correctly 57.02% of the stress classes. In order to find the general trend of confusion, the classification rate of individual listener is evaluated and it is shown in Figure 3.6. The individual bar plot depicts the variation in classification rate with different stress classes. Figure shows that most of the listeners confused the neutral speech with the sad speech and the angry speech with the Lombard speech. Sad speech is identified frequently. Further, Lombard speech is confused

Table 3.2: Performance of listener's stress classification for isolated keywords and segmented keyword in bracket

	Response Categories			
	Neutral	Angry	Sad	Lombard
Neutral	56.37 (56.10)	6.66 (10.01)	26.27 (21.25)	10.71 (11.83)
Angry	15.61 (10.23)	62.08 (60.71)	2.67 (3.34)	19.45 (25.23)
Sad	12.78 (15.86)	2.07 (2.39)	81.67 (79.36)	3.49 (2.39)
Lombard	29.60 (34.42)	28.79 (32.24)	6.97 (7.90)	34.64 (25.44)

either with the neutral or the angry speech.

Table 3.3: Average performance of listener's stress classification

	Average of keyword and segmented word	Good listeners	Good listeners and speakers
Neutral	56.10	58.41	54.98
Angry	61.41	71.46	74.90
Sad	80.51	84.30	88.95
Lombard	30.04	31.52	33.56
Average performance	57.02	61.42	63.10

Figure 3.6 also shows that some of the listeners are unable to distinguish among distinct stress classes such as listener 4, 9 and 14. These listeners are not able to discriminate between distinct stresses such as between sad and Lombard speech, and between neutral and angry speech. The average stress classification rate is also computed for each listener. It is observed that the average stress classification of these listeners are below to 50% of average stress classification rate of all listeners. Hence, these listeners can be removed from the listening test. In order to prune out such listeners, 50% of the individual average correct classification is considered as threshold. Only those listeners are chosen for stress classification who has more than 50% of average correct classification performance. The correct stress classification of listeners with this threshold is tabulated in Table 3.3. After pruning these bad listeners, the correct stress classification of listeners is 61.42%. This observation infers that all the listeners are not equally good at stress classification and some listeners are indeed bad. It may therefore be

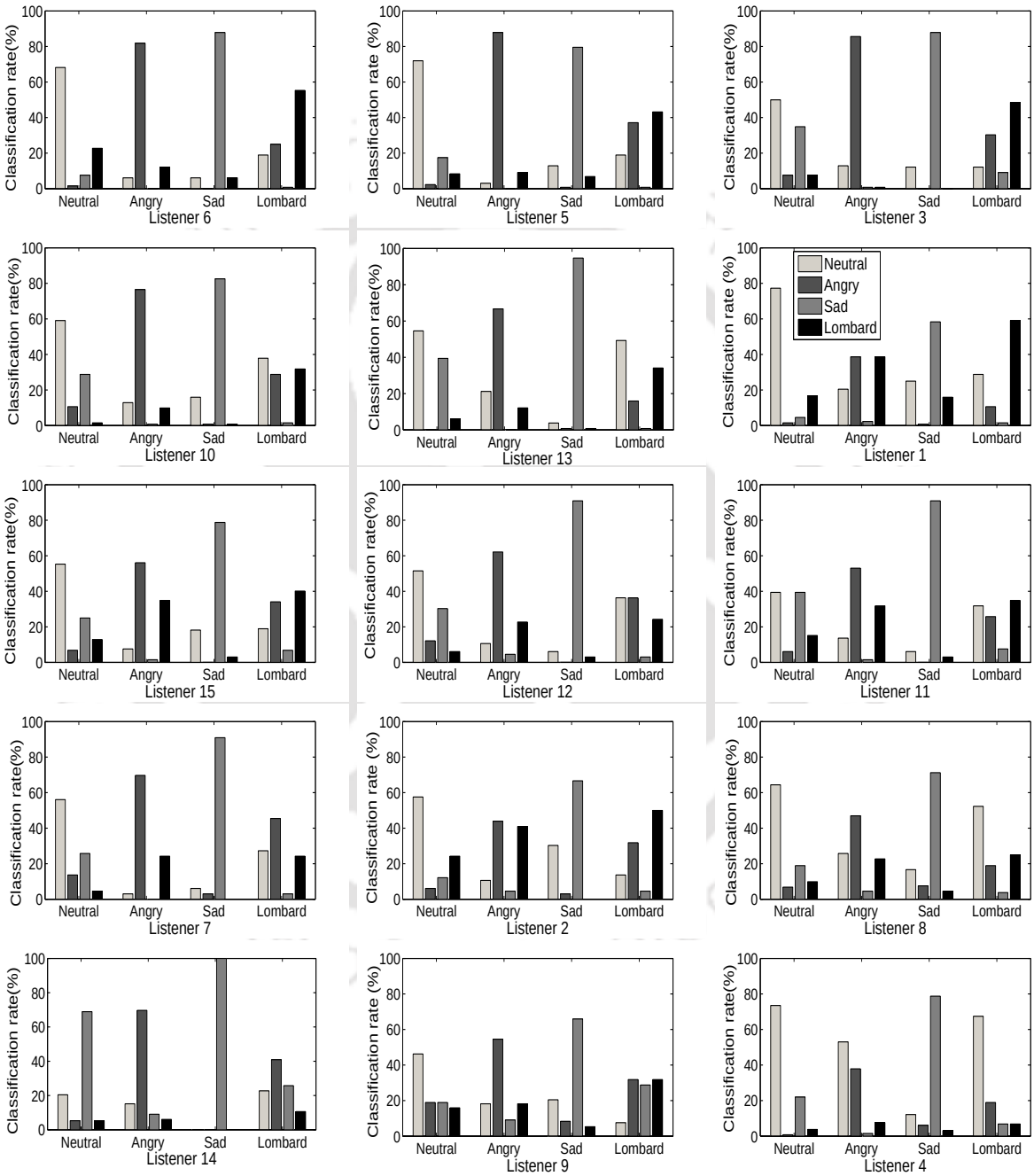

Figure 3.6: Classification rate of stress classification of individual listener

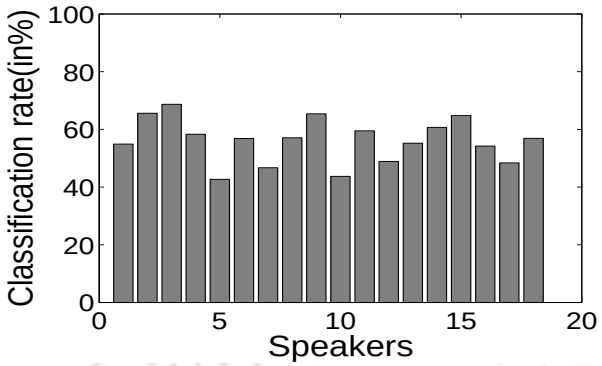

Figure 3.7: Average stress classification rate of individual speaker

necessary to eliminate the scores of such listeners.

Table 3.4: Performance of automatic stress classification using VQ

		Training			
		Neutral	Angry	Sad	Lombard
Testing	Neutral	49.64	11.73	19.31	16.28
	Angry	19.31	60.21	2.64	17.79
	Sad	26.50	8.32	61.36	3.78
	Lombard	15.52	23.10	1.89	59.46

Unlike listeners, some of the speakers in non-professional scenario are not able to produce stressed speech. The listening test sets contain seven utterances from each speaker for listening test. The classification rate of individual speaker depends on the correct classification of stress out of seven utterances from the fifteen listeners. The average classification rate for individual speaker is shown in Figure 3.7. Figure shows that speakers 5, 7, and 10 are not able to produce stressed speech of average standard. The average stress classification of these speakers are below 50% of average stress classification rate of all speakers. Hence, these speakers can be removed from the database. In order to prune out such speakers, 50% of the individual average correct classification is considered as threshold. Only those speakers are considered who has more than 50% of average correct classification. After removing these bad speakers from the database, the stress classification of speaker is tabulated in Table 3.3. Now, the stress classification of

listeners is 63.10%. This result infers that all the speakers are not able to produce stressed speech effectively. Therefore, listeners may not be able to identify the stress properly.

3.2.2.2 Automatic Stress Classification

This section presents the evaluation of stress level in the utterance of the two test sets by using automatic stress classification. The purpose is to demonstrate the ability of automatic methods to identify the stress present in an utterance. In this study, an automatic stress classifier is developed. Thirteen dimensional mel frequency cepstral coefficient ($MFCC$) (excluding c_0) is considered as a feature for stress classification [4]. The $MFCC$ feature of each frame is computed using frame size of 20 msec and frame rate of 10 msec. $MFCC$ features are extracted from approximately 12000 frames per stressed condition. The stress model is developed using VQ with codebook size of 1024 and by using continuous density HMM of sixteen states, left to right without skipping [30]. Each state has two mixture components. The performances of automatic stress classification using VQ and HMM techniques are tabulated in Tables 3.4 and 3.5, respectively.

Table 3.5: Performance of automatic stress classification using HMM

		Training			
		Neutral	Angry	Sad	Lombard
Testing	Neutral	41.66	6.81	37.12	14.39
	Angry	13.25	55.67	8.70	22.34
	Sad	10.98	6.43	77.64	4.92
	Lombard	24.99	23.47	8.32	43.17

The average performances of automatic stress classification using VQ and HMM are 57.66% and 54.53%, respectively. After removing bad speakers from the database, as mentioned in Section 3.2.2.1, the performances of automatic stress classification using VQ and HMM are tabulated in Tables 3.3. The average performances of automatic stress classification using VQ and HMM are 60.81% and 59.43%, respectively. These results infer that in non-professional scenario, some speakers are not able to produce stressed speech effectively, due to which neither listeners nor stress classifier method can identify the stress class from their speech. Hence, any

stress specific analysis on such speaker's speech may not give reliable information.

Table 3.6: Average performance of automatic stress classification

	VQ	HMM	Good speakers (VQ)	Good speakers (HMM)
Neutral	49.64	41.66	52.95	43.51
Angry	60.21	55.67	65.90	64.50
Sad	61.36	77.64	66.50	82.50
Lombard	59.46	43.17	57.87	47.21
Average performance	57.66	54.53	60.81	59.43

This section evaluated the level of stress present in the database when speech is recorded from non-professional speakers. From listening test, it was observed that listeners are able to perceive correctly 63.10% of the stress classes, whereas, automatic stress classification using VQ and HMM are able to identify 60.81% and 59.43%, respectively, correct stress classes. The stress identification rates of listener and automatic methods of this database are approximately similar to the most widely used $SUSAS$ database, as described in the Section 2.1. This observation infers that $SUSSC$ database also contains approximately same stress information as other stressed speech databases.

3.2.3 Evaluation of Speech Content

Section 3.2.2 demonstrates that speech signal contains stress information, which is validated by listeners and automatic stress classification method. The present section investigates the speech information in the database. To evaluate this, perceptual evaluation of speech content and automatic speech recognizer are used.

3.2.3.1 Human Perceptual Evaluation

In this study, a listening test has been done to evaluate the message of speech under different stressed conditions. Two sets are used for the listening test. Description of these two sets is same as given in Subsection 3.2.2.1. In this study, listeners are asked to identify the content of the speech under different stressed conditions. The same fifteen listeners are used for this

experiment. Listeners listened to each file and judged the content of the file. They had to write their option regarding the content. In this experiment, alternatives and forced choices are not given to the listeners, since the words have sufficient perceptual distinction from each other. The speech is considered as correctly identified, if the option of that file is matched with the content of that labeled file as given in Table 3.1. The listener's response for speech identification of isolated and segmented keywords is averaged and it is tabulated in Table 3.7. It is observed that listeners perceive correctly 99.85% of the content of the speech under the neutral condition. Under the stressed condition, 99.54% of the content of the speech is perceived correctly. This observation infers that listeners perception is less affected by the stressed condition.

Table 3.7: Performance of listener's stressed speech recognition

	Listener's recognition
Neutral	99.85
Angry	98.77
Sad	100
Lombard	99.53
Average performance	99.54

3.2.3.2 Automatic Speech Recognition

Generally, the system is trained in controlled environment, where it is assumed that the speaker is completely stressfree. Therefore, neutral speech is considered to model a keyword for practical interest. Thirteen dimensional mel frequency cepstral coefficient ($MFCC$) (including c_0) is considered as feature for speech recognition [4]. During training, approximately 75 utterances per word are considered from neutral speech of 1^{st} session database. $MFCC$ features of approximately 7000 frames per word are considered. It is observed from the experiments that to model a word, 128 codebook size is suitable for VQ technique and ten states and left to right model are suitable to model a word for HMM. Each state has two mixture components. The performances of automatic stressed speech recognizer using VQ and HMM are tabulated in Table 3.8. It is observed that the average performances of the recognizer using VQ and

HMM are 57.20% and 73.49%, respectively. It is also observed that the speech recognition performances of neutral speech for VQ and HMM are comparatively very high compared to those of other stressed conditions. This result infers that the performance of the recognizer is degraded significantly under the stressed conditions. However, listeners have prior knowledge of stress during speech recognition. Hence, to make a fair comparison to the listeners, knowledge of stress is given to the automatic speech recognition system. The speech from different stressed conditions of 1^{st} session of database are considered to develop word model. During training, approximately 75 utterances $\times$ 4 stressed conditions per word are considered. Model is developed using VQ technique with size of each word being 512 codebook. The performances of automatic stressed speech recognizer using VQ and HMM, are tabulated in Table 3.9. It is observed that, for VQ based recognizer, the average performance of recognizer is 81.05% whereas, for HMM based recognizer, the average performance is 76.14%. This result infers that after providing prior knowledge of stress during the training, the recognition system performs comparatively better than the recognition system trained with the neutral speech alone. However, the performance of speech recognition under the stressed condition is still not at par with that of neutral speech. Also, the degradation in the recognition performances for different stress speech from the neutral speech are not the same. This study infers that the characteristics of speech vary differently under different stressed conditions. As a result, the features extracted from these speech may also not have same characteristics. Therefore, the relative degradation of stressed speech is not same for different stressed conditions.

Table 3.8: Performance of stressed speech recognition for neutral speech trained system

	VQ	HMM
Neutral	86.37	87.88
Angry	39.39	65.16
Sad	46.97	69.70
Lombard	56.06	71.22
Average performance	57.20	73.49

Table 3.9: Performance of stressed speech recognition in case of system trained with different stressed conditions

	VQ	HMM
Neutral	89.39	84.85
Angry	84.85	74.24
Sad	68.15	66.67
Lombard	81.82	78.79
Average performance	81.05	76.14

3.2.3.3 Comparison of Speech Recognition for Different Stressed Speech

In this section, a comparison is done between confusion in the recognition for neutral speech and different stressed conditions. In this study, graphical representation of confusion matrix is used to compare confusion in the recognition for neutral speech and different stressed speech [12]. It maps confusion matrix onto 255 gray levels. Here, gray level 255 corresponds to 100% performance. In confusion matrix, higher performance of the system refers to the darkness of the block. For better visual distinction, only 25% of gray levels are considered to display the performance of the system.

The confusion matrix of performances of stressed speech recognition under different stressed conditions in case of a system trained with different stressed conditions using VQ are shown in Figure 3.8. In this figure, x-axis depicts the word model and y-axis depicts the recognition performance of stressed speech tested with each word model. By comparing Figure 3.8(a) and Figures 3.8((b)-(d)), it is observed that, less off-diagonals are present in case of neutral speech compared to other stressed conditions. This observation infers that the confusion in neutral speech is mainly due to the inability of techniques employed for recognition. On the other hand, under different stressed conditions, off-diagonal elements are more which shows a considerable confusion due to stress. Hence, new techniques are required to compensate the effect of stress.

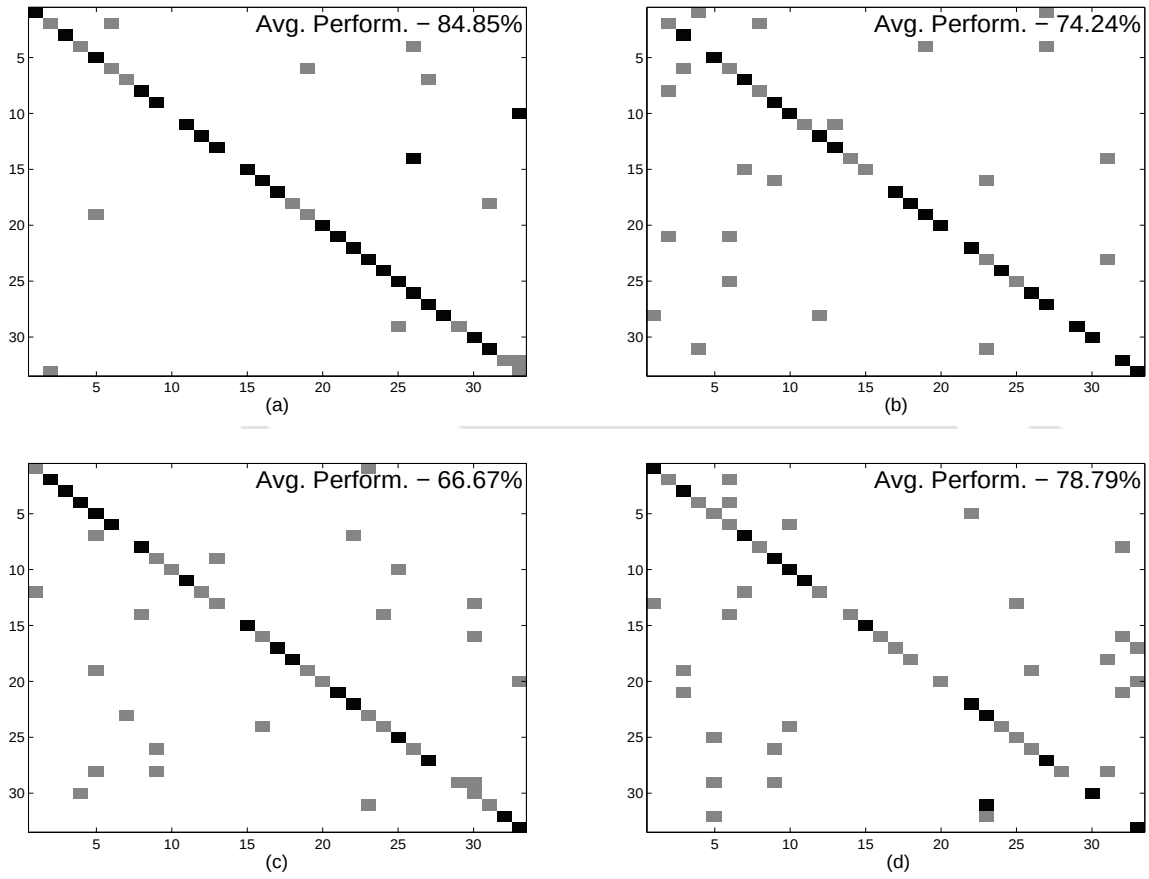

Figure 3.8: Confusion matrix of performance of speech recognition under different stressed conditions. Here, x-axis depicts the word model and y-axis depicts the recognition performance of stressed speech tested with each word model. (a) neutral condition, (b) angry condition, (c) sad condition and (d) Lombard condition

3.3 Performances of Speech Features for Stressed Speech Recognition

It is observed from the Section 3.2.3 that the speech recognizer gives poor performance when speech is stressed. This observation motivates us to evaluate the effectiveness of different speech features for speech recognition under different stressed conditions and to evaluate the relative degradation in the recognition performances using these features under different stressed conditions. In this section, the performances of mel frequency cesptral coefficient ($MFCC$) and linear predictive cepstral coefficient ($LPCC$) features along with their derivatives are re-evaluated under four stressed conditions [11] [10]. Two existing approaches of stressed speech recognition namely, multi-style training and adaptive cepstral mean normalization techniques are also re-evaluated.

Automatic speech recognition is developed for 119 word vocabulary. Speech of only those speakers are considered who have more than 50% average stress classification rate, as explained in Section 3.2.2.1. Thus, the speech from fifteen speakers is considered for further analysis. During the training, approximately 40 speech tokens per word of neutral speech are used from 1^{st} session of database. Thirteen dimensional features are extracted from approximately 5000 frames per word. Word model is developed using HMM classifier. Word model is developed using ten states and left to right model. Each state has two mixture components. For testing, approximately 40 speech tokens per word of different stress classes are used from 2^{nd} session of database. Table 3.10 gives the performances of automatic speech recognition using different features under different stressed condition.

First, the $MFCC$ feature is extracted from neutral speech to develop a word model during training. During testing, $MFCC$ feature is tested with speech under different stressed conditions along with neutral speech. The average performance of recognizer under the stressed condition is 58.40% and that for neutral speech is 79.22%. This result infers that the performance of the system degrades significantly under stressed condition. This is mainly due to the variation in the cepstral coefficient extracted from the stressed speech. In $MFCC$ feature, 10 to

Table 3.10: Performance of speech recognition using different features under different stressed conditions

Feature	Neutral	Angry	Sad	Lombard	Avg. Performance
$MFCC$	79.22	40.01	60.38	54.00	58.40
$MFCC$+Δ+ΔΔ	84.80	47.25	65.78	62.76	65.15
$MFCC$+Preemphasis	81.13	42.58	59.89	55.69	59.82
$MFCC$+Δ+ΔΔ +Preemphasis	84.87	52.15	66.87	65.61	67.38
LPCC+Preemphasis	63.79	44.83	46.38	40.46	48.86
LPCC+preemph+Δ+ΔΔ	79.41	63.04	63.47	64.29	67.55
Multi-style training	80.42	64.82	63.64	75.14	71.14
Adaptive CMN	73.54	39.86	53.43	52.14	54.74

15 initial cepstral coefficients are used which contains the slow varying log magnitude spectrum of the speech [44]. The spectrum under stressed condition varies compared to neutral condition. Due to this variation in spectrum, the cepstral coefficient of speech under stressed condition varies from that under neutral condition. Hence, the performance of the recognizer is degraded. The performance of the recognizer improves from 58.40% to 65.15% after applying Δ and ΔΔ in $MFCC$ feature, as described in Chapter 2. The phonetic information in speech may get encoded in its spectral change over time. Hence, it is expected that the dynamic and acceleration features will capture this information better as compared to static feature [95]. However, stress may modify spectral characteristic uniformly over a short period such as phonemes and syllables of speech. This change in spectral characteristics may affect the static features of speech but dynamic and acceleration features will remain unchanged under stressed condition [44, 95]. After applying preemphasis, as described in Chapter 2, the performance of recognizer improves from 58.40% to 59.82% and from 65.15% to 67.38% for $MFCC$ and $MFCC$ with Δ and ΔΔ, respectively. Preemphasis filter is used to emphasize the high frequency contents. Preemphasis filter reduces the dynamic range in the frequency domain by flattening of spectral tilt, which improves linear modeling of the formant structure [69]. In $MFCC$ computation, the spectrum of the speech signal contains both vocal tract and excitation source information. The degradation in the recognition performance is mainly attributed to the variation in slow varying spectrum,

which may contain vocal tract information. In order to investigate this assumption, the performance of $LPCC$ feature is evaluated under different stressed conditions. $LPCC$ feature is derived from the LP spectrum, which eliminates the excitation source information and contains only the vocal tract information. The thirteen dimension $LPCC$ feature from neutral speech is considered for training and testing. The trend of recognition performance of $LPCC$ feature is observed to be similar to the $MFCC$ feature. The average performance of recognizer is 48.86% and after applying Δ and $\Delta\Delta$ in $LPCC$ feature, the average performance of recognizer improves from 48.86% to 67.55%. These results further support that the degradation in the performance is mainly due to the variation in slow varying spectrum, attributed to the variation in vocal tract spectrum. In the next study, multi-style training approach is evaluated for stressed speech recognition. In this approach, stressed utterances under angry, sad and Lombard conditions along with the neutral speech are used to develop a word model during the training. During the training, approximately fifty utterances per stressed condition are taken from 1^{st} session of database. Thus, 40 utterances $\times$ 4 stress classes number of utterances are used to develop a word model. The average performance of this approach is given in Table 3.10. The average performance of recognizer is improved from 59.82% to 71.14%. Although, this technique gives sufficient improvement in the recognition performance, it requires additional stressed speech during the training, which is practically difficult. The adaptive mean cepstral compensation (CMN) based stress compensation technique is used for stressed speech recognition [11]. This technique assumed stress component as additive to the speech component. Performance of this approach is also evaluated for $SUSSC$ database and tabulated in Table 3.10. It is observed that the performance of this technique degrades from the baseline.

3.4 Summary

This chapter demonstrates the development and evaluation of a stressed speech database. In the study, the stress information is evaluated using listening test and automatic methods for classification of stress classes. This study showed that perceptual and automatic stress classifier performs approximately similar stress classification as the $SUSAS$ database. From this study,

it can be inferred that the analysis of stress using $SUSSC$ database may give approximately similar stress information as other stressed speech database. The next study evaluated speech information using listening test and automatic stressed speech recognition. It is observed that the listener's perception of speech is not effected by stress. For automatic speech recognition, performance of the system trained with neutral speech degrades more than that of a system trained with different stressed conditions. Compared to neutral speech, confusion is observed to be more under the stressed conditions. The $MFCC$ and $LPCC$ features with their derivatives are evaluated for stressed speech recognition performances. The performances of these features under stressed condition degrade as compared to those of neutral speech. The degradation in the performance is mainly attributed to the variation in the slow varying vocal tract spectrum. Hence, there is need to explore robust techniques to improve the performance of the system.

4

Analysis of Relative Displacement of Formant Peaks

Contents

It was observed in Chapter 3 that the conventional mel-frequency cepstral coefficient ($MFCC$) and linear predictive cepstral coefficient ($LPCC$) features provide good recognition performances under stressfree or neutral environment. However, these features give poor performances when speech is stressed. Chen [11] observed that the degradation in the performance is mainly due to the average shift of the mean values and average scaling of the variances of the cepstral coefficients. The shift in mean values of cepstral coefficients was observed due to the variation in slope of the spectrum (spectral tilt). Hansen [81] observed that the spectral tilt depends on the glottal parameters. The glottal parameters are opening slope, closing slope, opening duration, top duration, closing duration, and closed duration [79, 81]. The slopes of glottal opening and closing provide control over the amount of acoustic energies produced. The duration of glottal closure decides the length of pitch period. Under certain stressed conditions, such as angry, loud and Lombard, glottal pulses have sharp rise times and sharp corners due to variations in the glottal parameters. As a result, in the spectrum, the energy of higher frequency is relatively increased from that of lower frequency [79]. Therefore, slope of the spectrum becomes more flat under these stressed conditions. Variation in the glottal vibration effects not only the migration of spectral energy but also the vocal loudness. Seshadri and Yegnanarayana [13] observed abruptness in the closing phase of glottal cycle under angry, loud and Lombard conditions, which produces variation in the perception of loudness. The variation in the loudness produces variation in the formant peaks (amplitudes). Klatt [83] observed that the change in spectral tilt and bandwidths of the spectrum do not produce any change in the perception of vowel identity, but it gives information about variation in the speaker. Lu and Cook [84] observed that the intelligibility and the loudness of the speech are increased due to increase in flatness of spectral slope for Lombard speech. From all these observations, it can be concluded that the stress effects the slope or tilt of the spectrum.

This chapter focuses on analysis of stress in the spectrum using spectral tilt as a feature. Tartter et al. [17] considered spectral tilt as a measure for the distribution of spectral energy from the lower formants to the higher formants. The spectral tilt was measured as a slope of a regression line fit to the short-term power spectrum of high intensity vowel region. This

slope gives gross energy variation across the spectrum [17, 18]. Koolagudi and Krothapalli [15] defined spectral tilt as a positive angle between the line joining of 1^{st} formant peak and 2^{nd} formant peak with an abscissa. This study showed the effect of stress in the first two formant peaks. However, all formant peaks may get effected due to the migration of energy from the lower formants to the higher formants. Hence, it can be inferred that the variation in the spectral tilt is not only due to migration of spectral energy, but also due to the variation in formant peaks, which is introduced as a consequence of migration of spectral energy as well as loudness. In this chapter, the effect of stress on the formant peaks is studied by proposing local spectral tilt as a measure. Local spectral tilt refers to the relative variation between formant peaks and it is named as Relative Formant Peak Displacement (RFD). The effect of stress is investigated using the local and the gross spectral tilts and they together form a feature termed as RFD feature. The stress information in this feature is evaluated by considering this feature of classification of stressed speech.

Analysis of gross spectral tilt under different stressed conditions is studied in Section 4.1. The analysis of stress using the proposed RFD feature is studied in Section 4.2. Section 4.3 presents the evaluation of stress classification using RFD feature. Section 4.4 describes stressed speech recognition using proposed stress classifier. Section 4.5 gives summary of this chapter.

4.1 Analysis of Gross Spectral Tilt of Speech Under Stressed Condition

The migration of spectral energy depends on the glottal parameters. These parameters may not be equal for different stressed conditions, due to which, the migration of spectral energy would be different under different stressed conditions. In the present section, the migrations of spectral energy under different stressed conditions are analyzed using gross spectral tilt as a measure. In the present work, gross spectral tilt is considered as a measure to find the spectral energy distribution of speech. Spectral tilt can be defined as the relative distribution of spectral energy from the lower frequency to the higher frequency [17]. A linear regression line is fitted to the spectrum of the segment of vowel region using least square error method [17, 18]. The

slope of the regression line is taken as a measure of spectral tilt.

The effect of stress on spectral tilt is analyzed by using four frames, namely, two frames of neutral speech and other two frames of angry speech. In this analysis, 20 msec segment of vowel region of speech is considered as a frame. These frames are taken from the same speaker to avoid speaker variability. For each frame, a 256-point log magnitude spectrum is computed using 10^{th} order LP coefficient of that frame. A regression line is fitted to that spectrum and the slope of the line is taken as a spectral tilt of that frame. The log magnitude spectra of these frames and their spectral tilts for vowels /a/ are shown in Figure 5.1. For each stress, frames from preemphasized speech and non-preemphasized speech are considered for comparison. Figure 5.1 (a),(b) and Figure 5.1 (c),(d) represent spectra of two frames of non-preemphasized neutral speech and two frames of non-preemphasized angry speech, respectively. It is observed that the spectral energy of angry speech at higher frequencies is more, which increases the flatness of the spectral tilt compared to that of neutral speech. Angry speech has the glottal pulse with sharp rise time and sharp corner, due to which the high frequency content may increase [96]. As a result, the flatness of tilt of the spectrum increases. Figure 5.1 (e),(f) and Figure 5.1 (g),(h) represent spectra of two frames of preemphasized neutral speech and two frames of preemphasized angry speech, respectively. It is observed from these figures that the energy at higher frequency are enhanced, which reduces the dynamic range in the frequency domain of the speech signal. Due to this, the slope of spectrum of preemphasized speech become flatter compared to non-preemphasized speech.

The flatness of spectral tilt under angry condition infers that the spectral tilt contains stress information of the speech signal. In order to evaluate the relative variation of spectral tilt under different stressed conditions, PDF based statistical evaluation is used. The spectral tilt of vowels, $/a/$, $/e/$, $/i/$, and $/u/$ under different stressed conditions are considered for this analysis. In order to estimate the PDF of the spectral tilt, approximately 250 stressed frames of a vowel per stressed condition are used for this analysis. Here, stressed frames refer to those frames where the intensity of vowel is high. As described in Chapter 2, preemphasis filter suppresses glottal source and radiation effect from the speech signal, due to which, the

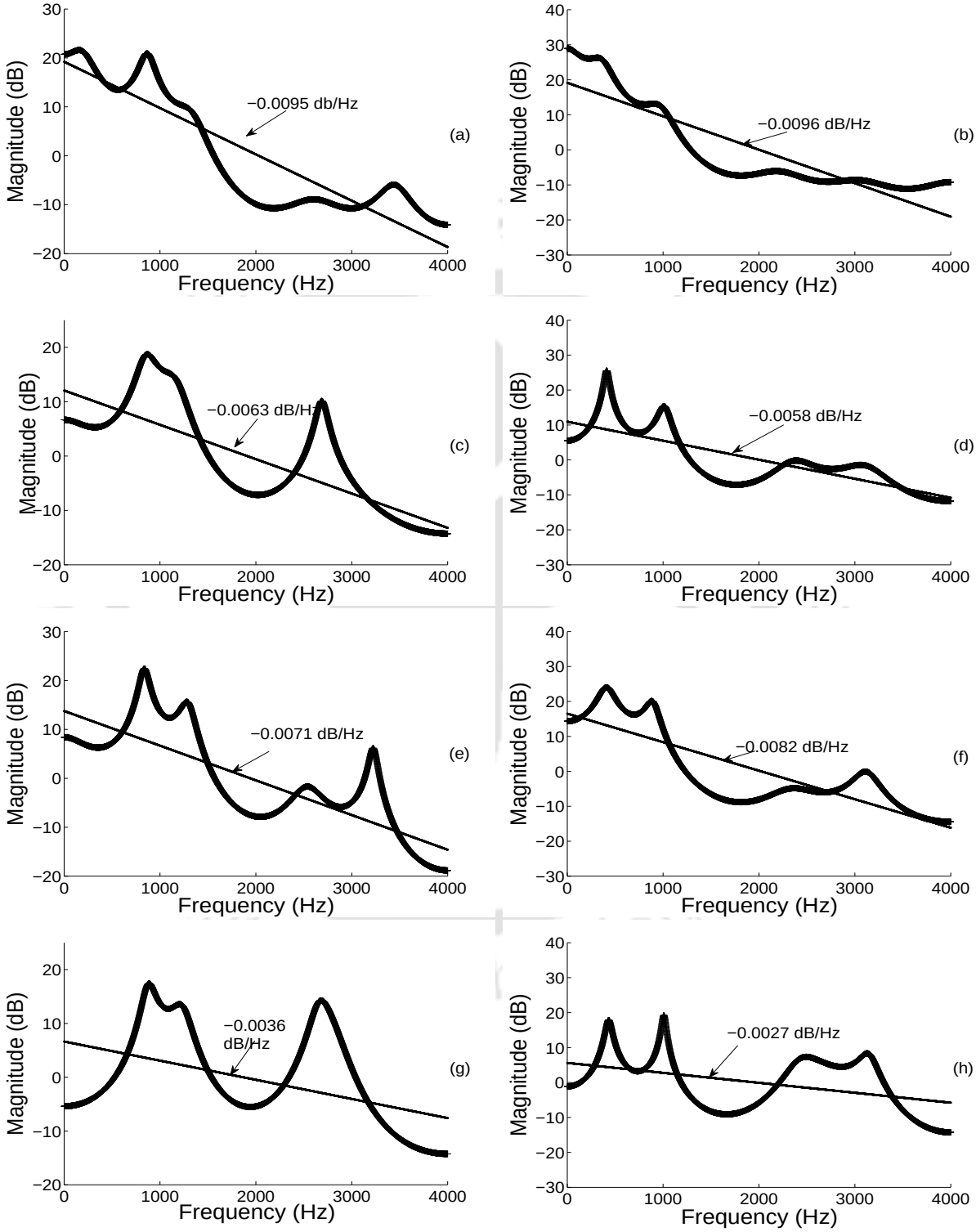

Figure 4.1: Spectrum and its spectral tilt of vowel /*a*/ for (a),(b) neutral speech without using preemphasis, (c),(d) angry speech without using preemphasis, (e),(f) neutral speech using preemphasis, and (g),(h) angry speech using preemphasis

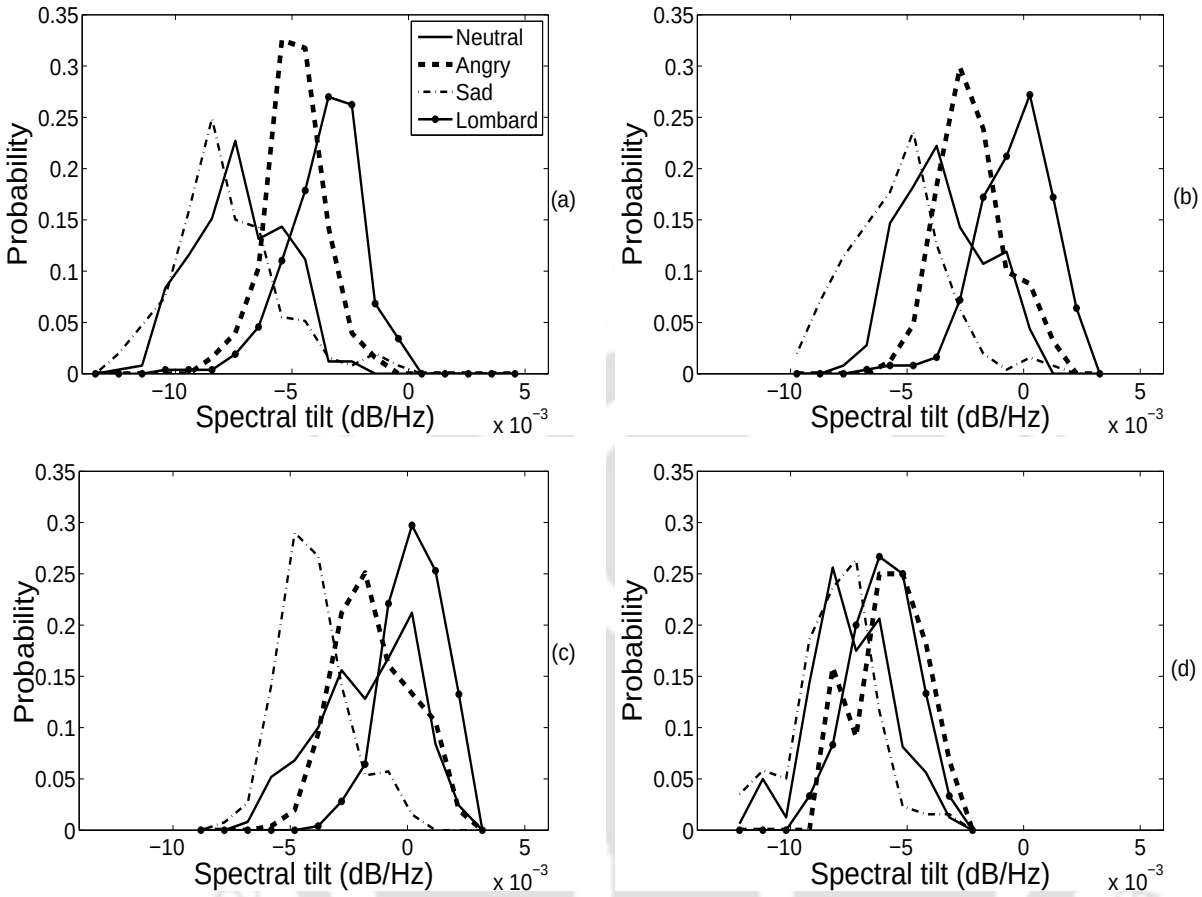

Figure 4.2: Distributions of spectral tilt of vowels under different stressed conditions for (a) /*a*/, (b) /*e*/, (c) /*i*/ and (d) /*u*/

formant structure is better estimated. Therefore, in the present study, preemphasized speech is considered for analysis. The maximum and minimum values of spectral tilt are determined and the difference between them are divided equally to find the bins of the distribution. The PDF of spectral tilt of a vowel is then estimated by the histogram method. Figure 4.2 shows the distributions of spectral tilt of four vowels under different stressed conditions. Figures show that the peaks of the distributions of spectral tilt under stressed conditions are not at the same place. The mean value of spectral tilt varies under stressed conditions for a vowel. Also, the spectral tilt under different stressed conditions have non-overlapping PDF characteristics. This shows that the distributions of spectral tilts under stressed conditions are quite separated from each other. These observations are made on the basis of visual interpretation of the

PDF characteristics. In order to quantify these differences, statistical evaluations such as mean, variance and divergence values of the spectral tilt are estimated under different stressed conditions. The mean value of spectral tilt of a vowel under a given stressed condition is computed as the average of spectral tilt of 250 stressed frames of that vowel. The mean value of spectral tilt of neutral, angry, sad and Lombard speech are -4.9, -3.4, -6.2 and -2.1 (in 10^{-3}), respectively. It is inferred that the mean values of spectral tilt for angry and Lombard speech are relatively higher and the mean value of spectral tilt for sad speech is lower than that of neutral speech. From these observations, it is inferred that the mean of spectral tilt varies significantly under stressed conditions from that of neutral condition. The variance values of spectral tilt under different stressed conditions are computed in similar way as described for mean computation of spectral tilt. The variance of spectral tilt of neutral, angry, sad and Lombard speech are 3.38, 2, 3.11 and 1.7 (in 10^{-6}), respectively. It is observed that, the variance values of spectral tilt of angry and Lombard speech are less compared to those of neutral and sad speech. These results infer that the variance of spectral tilt also varies under stressed conditions. The distribution is uniquely defined by its mean and variance values. From mean and variance of spectral tilt based analysis, it is observed that the mean and variance of spectral tilt vary under stressed conditions.

To quantify the deviation in the distribution of tilt of stressed speech from that of neutral speech, Kullback-Leibler divergence (KLD) is used [97]. KLD is defined as a relative entropy between two density functions, $f(x)$ and $g(x)$. The KLD is calculated by Eq 4.1, where, $x_i, i = 1...N$ are the N sample from $f(x)$ and $g(x)$, respectively. The relative entropy between two distributions is null if the two distributions are identical. Thus, the divergence between two distributions determines how distinct the distribution $f(x)$ is from the distribution $g(x)$. In this study, the KLD of distribution of spectral tilt of stressed speech $g(x)$ from that of neutral speech, $f(x)$ determines how distinguishable the distribution, $f(x)$ is from the distribution $g(x)$. KLD between $f(x)$ and $g(x)$ is given in

$$\mathrm{D_{kl}}(f||g) = \sum_{i=1}^{N} f(x_i) \log \frac{f(x_i)}{g(x_i)} \tag{4.1}$$

where, $f(x_i)$ and $g(x_i)$ represent the distributions of spectral tilt of neutral and stressed speech, respectively at i^{th} sample. The divergence of spectral tilt of a vowel under a given stressed condition is computed using distribution of spectral tilt formed by 250 stressed frames of the given vowel. The average divergence of spectral tilt of stressed speech is computed by considering average of divergences of spectral tilt of vowels. The divergences of angry, sad and Lombard speech from the neutral speech are 6.06, 2.15 and 9.67, respectively. It is observed that the deviations of distributions of spectral tilt for angry and Lombard speech from neutral speech are comparatively higher than that of sad speech. This observation infers that the spectral tilt of angry and Lombard speech are more flat. It shows that more spectral energy are shifted to higher frequency under angry and Lombard conditions compared to the sad condition. This means that the migration of spectral energy for sad speech is comparatively lesser due to stress than those of angry and Lombard speech. For sad speech, the glottal closing period is more as compared to the neutral speech, whereas, the glottal closing period is less for angry and Lombard speech [81]. This may be the reason for migration of less spectral energy for sad speech and migration of more spectral energy for angry and Lombard speech. This observation also infers that the distributions of spectral tilt under different stressed conditions are significantly different from each other and hence, the spectral tilt has ability to discriminate stress.

4.2 Proposed Relative Formant Peak Displacement for Quantification of Local Spectral Tilt

In the previous section, the effect of stress is analyzed on the gross spectral tilt and it is found that the spectral tilt has significant discriminating capability for stress. Since, the spectral tilt includes gross energy information of the spectrum, it may lose the variation in formant peaks induced due to the loudness of speech. However, the loudness of the signal varies due to variation in glottal closure. This variation is directly reflected in the formant peaks. It can be said that the formant peak is a cue factor of stress. Hence, the quantification of this variation may give stress information. In order to capture the variation in formant peaks due to stress, spectral tilt is divided into the local spectral tilts. The local spectral tilt can be obtained by

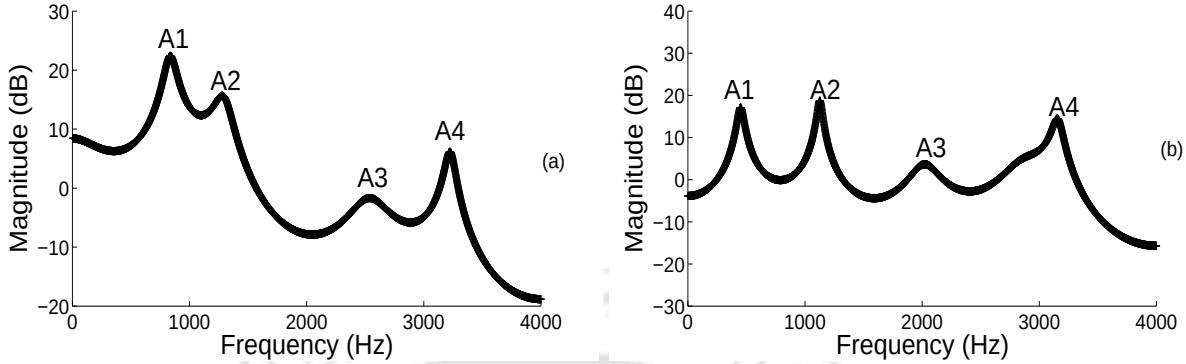

Figure 4.3: Spectrum of vowel /*u*/ for two stressed conditions (a) neutral and (b) angry

considering the relative variations in formant peaks.

Figure 4.3 shows LP spectra of a segment of vowel /*u*/ of neutral and angry speech. From Figure 4.3 (a) and (b), it is observed that the amplitude deviation from 1^{st} formant peak (A_1) to 2^{nd} formant peak (A_2) under angry condition is less than that under neutral condition. Similarly, the amplitude deviations from 1^{st} formant peak (A_1) to 3^{rd} formant peak (A_3) and 4^{th} formant peak (A_4) are less than those under the neutral condition. Due to the reduction in amplitude deviation of A_2, A_3 and A_4 from A_1 in angry speech, the flatness of spectral tilt increases with respect to the neutral speech. This observation shows that relative deviations of formant peaks from 1^{st} formant peak may contain stress specific information. The relative deviation in formant peaks from 1^{st} formant peak is defined as *Relative Formant Peak Displacement* (*RFD*). The *RFD* of i^{th} formant peak (A_i) is defined as the relative displacement of the i^{th} peak from first formant peak (A_1) as given in Eq. 4.2.

$$RFD_i = \frac{A_1 - A_i}{|A_1|} \qquad 2 \leq i \leq 4 \tag{4.2}$$

The *RFD* values of formant peaks have approximately inverse relation with flatness of spectral tilt. Less *RFD* value for a formant peak will correspond to more flatness in spectral tilt for that formant peak. For *RFD* computation, the formant peaks can be obtained from *LP* spectrum and cepstrally smoothed log spectrum [2]. In *LP* spectrum, the formant peaks are highly influenced by the artifacts of *LP* analysis [2]. The shift in formant frequency and

the change in formant bandwidth directly influences the intensity of the peak. Therefore, the spectral peak measurement may not be reliable for LP based analysis. On the other hand, cepstrally smoothed log spectrum removes the sensitivity of spectral intensity without altering the formant structure [2]. In this section, RFD values are estimated from both LP spectrum and cepstrally smoothed spectrum for comparison. The block diagram for extraction of RFD

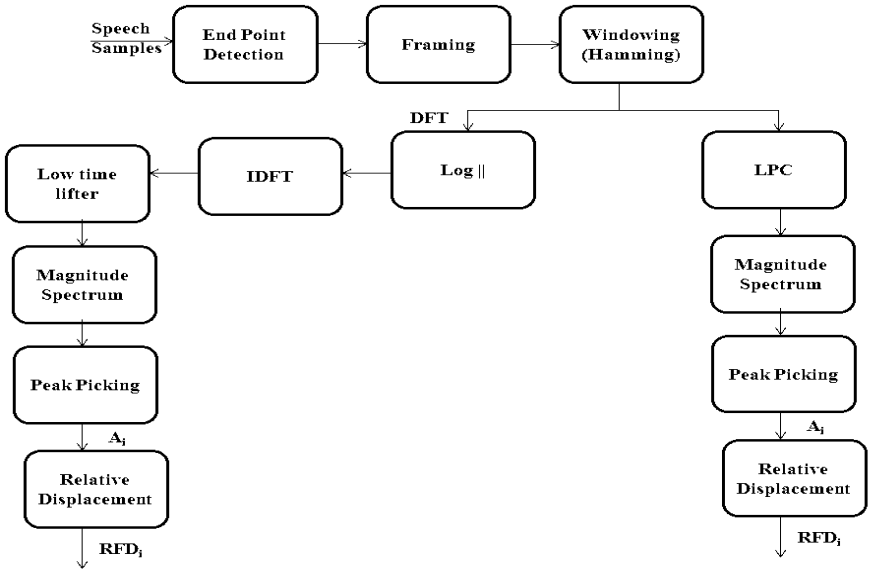

Figure 4.4: Block diagram of extraction of relative formant peak displacement (A_i - Formant peak)

values, derived from LP spectrum and cpestrally smoothed log spectra, are shown in Figure 4.4. The speech signal is segmented into a number of frames with length of 160 samples and frame shift of 80 samples. To remove the silence regions from the speech signal, end point detection is used [2]. Only those frames are considered as speech frames whose energy is higher than 10% of average energy of the signal. Hamming window is applied on those frames to reduce the discontinuities at the end of the frames. Ten LP coefficients are estimated from each frame and the magnitude spectrum is computed from these LP coefficients. The formant peaks A_i's are

Table 4.1: Means and variances of RFD values for different stress classes for vowels

RFD_2	μ_N	σ_N^2	μ_A	σ_A^2	μ_S	σ_S^2	μ_L	σ_L^2
/a/	0.4004	0.0239	0.3247	0.0102	0.1888	0.0331	0.0812	0.0834
/e/	0.7185	0.7185	0.2614	0.3593	0.8478	0.0372	-0.2259	0.6679
/i/	0.6139	0.6139	-0.3037	0.4037	0.5319	0.0271	0.295	0.0025
/u/	0.2024	0.0075	0.019	0.0166	0.3682	0.0428	0.0243	0.016
RFD_3	μ_N	σ_N^2	μ_A	σ_A^2	μ_S	σ_S^2	μ_L	σ_L^2
/a/	0.7369	0.0221	0.4979	0.0087	0.6092	0.0552	0.1815	0.1123
/e/	0.6843	0.0287	-0.0099	0.1815	0.7345	0.0239	-0.0441	0.6899
/i/	0.8475	0.0136	-0.2648	0.1649	0.3769	0.0691	0.1405	0.0114
/u/	1.1341	0.0103	0.6321	0.0568	1.1276	0.0011	0.6469	0.0347
RFD_4	μ_N	σ_N^2	μ_A	σ_A^2	μ_S	σ_S^2	μ_L	σ_L^2
/a/	1.2688	0.026	0.7162	0.0657	0.8843	0.0479	0.4933	0.0306
/e/	0.7044	0.0191	0.4052	0.0644	0.6953	0.0439	-0.1597	0.5131
/i/	0.5773	0.0257	-0.0547	0.111	0.6913	0.0721	0.5258	0.0223
/u/	1.0947	0.01	0.5972	0.0059	1.1242	0.0024	0.5215	0.0315

extracted from the spectrum using peak picking algorithm. Relative formant peak displacement of 2^{nd}, 3^{rd} and 4^{th} formant peaks are computed using Eq. 4.2 and these values are named as RFD_2, RFD_3 and RFD_4, respectively. The RFD values reflect the relative variation between formant peaks. In order to evaluate the stress information from these values, probability density function, as described in Section 4.1, is considered. These RFD values are computed for all stressed frames under different stressed conditions. Approximately, 5000 spectra of vowel /a/ are considered under stressed condition. It gives approximately 5000 RFD values of vowel /a/. The distribution of RFD_2, RFD_3 and RFD_4 under different stressed conditions are shown in Figure 4.5. Figure shows that the means and variances of RFD values vary under stressed conditions. Also, the distributions are observed significantly different from each other. From the figure, it can be observed that the mean of RFD_4 is higher than the mean of RFD_3, and the mean of RFD_3 is higher than the mean of RFD_2. The distributions of RFD values vary with stressed condition. The deviation of distribution of RFD_4 value under different stressed conditions are more than those of RFD_3 and RFD_2 values. This observation shows that the effect of stress is more at higher frequency than at lower frequency.

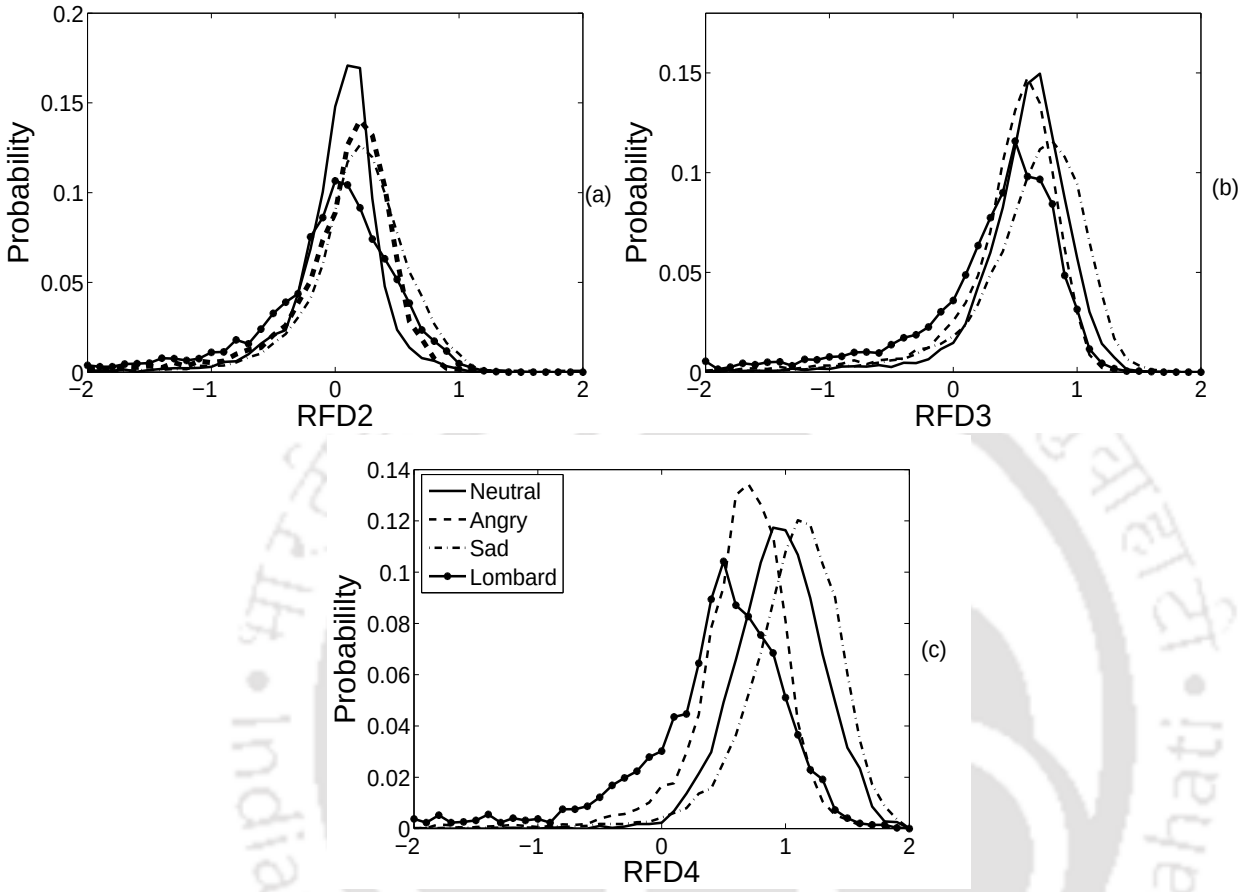

Figure 4.5: Distributions of RFD values for vowel /a/ under different stressed conditions (a) RFD_2 (b) RFD_3 and (c) RFD_4

The mean and variance values of RFD_2, RFD_3 and RFD_4 for vowels, /a/, /e/, /i/, and /u/ under different stressed conditions are listed in Table 4.1. μ_N, μ_A, μ_S, μ_L and σ_N, σ_A, σ_S, σ_L are the means and variances of RFD values for neutral, angry, sad and Lombard speech, respectively. From this table, it is observed the mean and variance of RFD_2, RFD_3 and RFD_4 values under angry, sad and Lombard conditions are less than that under neutral condition. Under angry, sad and Lombard conditions, RFD_3 and RFD_4 values are less compared to RFD_2 values. However, RFD_2, RFD_3 and RFD_4 values under the sad conditions are higher than those under angry and Lombard conditions. These observations infer that RFD_3 and RFD_4 values are less under angry and Lombard conditions, due to which, the flatness of spectral tilt of speech increases under these conditions. On the other hand, RFD_3 and RFD_4 values are

Table 4.2: Divergences of RFD values for different stress classes for vowels

	Angry			Sad			Lombard		
	RFD_2	RFD_3	RFD_4	RFD_2	RFD_3	RFD_4	RFD_2	RFD_3	RFD_4
/a/	3.07	15.98	37.52	2.73	6.58	25.80	5.73	20.09	38.51
/e/	19.47	24.73	36.37	17.86	13.92	29.90	21.80	22.20	34.06
/i/	11.21	13.11	23.91	5.77	4.86	12.90	11.43	11.28	20.04
/u/	28.28	37.44	43.62	29.56	32.44	41.42	33.33	41.72	44.27

higher under sad condition, due to which, the flatness of spectral tilt of speech reduces under this condition.

Due to variations in the means and variances of RFD values, the distributions of RFD values under the stressed condition are separated from those of neutral condition. In order to quantify this deviation, KL divergence between neutral speech and stressed speech is computed using Eq. 4.1. The divergence of RFD_2, RFD_3 and RFD_4 values of vowels, /a/, /e/, /i/, and /u/ under angry, sad and Lombard conditions are tabulated in Table 4.2. From this table, it can be observed that the divergence of RFD_3 and RFD_4 for stressed speech are more than the divergence of RFD_2. From all these observations, it can be concluded that formant peaks A_3 and A_4 are more effected due to stress compared to A_2. Further, the divergences of RFD values for sad speech are less than those for angry and Lombard speech. This observation infers that the formant peaks of sad speech are comparatively less effected from those of angry and Lombard speech. As described in Section 4.1, the glottal closing period is more in case of sad speech and it is less for angry and Lombard speech. The loudness of angry and Lombard speech are more and it is less for sad speech. From these observations, it is also inferred that the RFD values have discriminating capability for stress. Hence, these RFD values are expected to contain stress information and they can be used to characterize the stress.

4.2.1 Average Stress Relative Formant Peak Displacement

Previous subsection considered relative deviations of formant peaks from 1^{st} formant peak. These values are derived within the spectrum for a given stress class, therefore, it gives intra

formant peak deviation information. Intra formant peak deviation information is a relative information from a spectrum. This section analyzes the inter formant peak deviation information. It is described previously in the current section that the formant peaks vary under stressed conditions due to variation in loudness of speech. Also, it can be observed from Figure 4.3 (a) and (b) that the difference between amplitude values of 1^{st} formant peak of neutral speech and angry speech is approximately 5 dB (25dB-20dB). On the other hand, difference between amplitude values of 2^{nd} formant peak of neutral speech and angry speech is approximately -7 dB (18dB-23dB). *Stress Relative Formant Displacement* ($SRFD$) is proposed in order to quantify the degree of deviation of formant peak of stressed speech from that of neutral speech. Stress relative formant displacement is defined in Eq. 4.3, where, A_{iN} and A_{iS} denote formant peaks of neutral and stressed speech, respectively.

$$SRFD_i = \frac{A_{iN} - A_{iX}}{|A_{iN}|} \qquad 1 \leq i \leq 4 \tag{4.3}$$

$$ASRFD_i = \frac{1}{M}\sum_{k=1}^{M}\frac{A_{ikN} - A_{ikX}}{|A_{ikN}|} \qquad 1 \leq i \leq 4 \tag{4.4}$$

M represents the number of frames for a stress class, i denotes formant peaks including 1^{st} formant peak and X denotes the stress classes (excluding neutral class).

The $SRFD$ values quantify the relative deviation of formant peaks of stressed speech from the corresponding formant peaks of neutral speech. In order to investigate the deviation in the formant peaks due to stress, $SRFD$ values are computed for vowels, /a/, /e/, /i/, and /u/ under neutral, angry, sad and Lombard conditions. Approximately 5000 frames under stressed condition are considered. $SRFD$ values are estimated for vowels for angry, sad and Lombard speech using Eq. 4.3. Further, the *Average Stressed Relative Formant Peak Displacement* is further computed by averaging of $SRFD$ values of frames for i^{th} peak. The average stress relative formant displacement of four vowels under different stressed conditions are shown in Figure 4.6. From Figure 4.6(a), it can be observed that the displacement of 1^{st} formant peak for stressed speech is not as significant as the other three formant peaks. These observations are also observed for all vowels. The displacement for angry and Lombard speech is negative.

Hence, the formant peaks for angry and Lombard speech are higher than the formant peaks of neutral speech. On the other hand, the displacement of sad speech is positive for most of the peaks. Hence, the formant peaks of sad speech are smaller than the formant peaks of neutral speech.

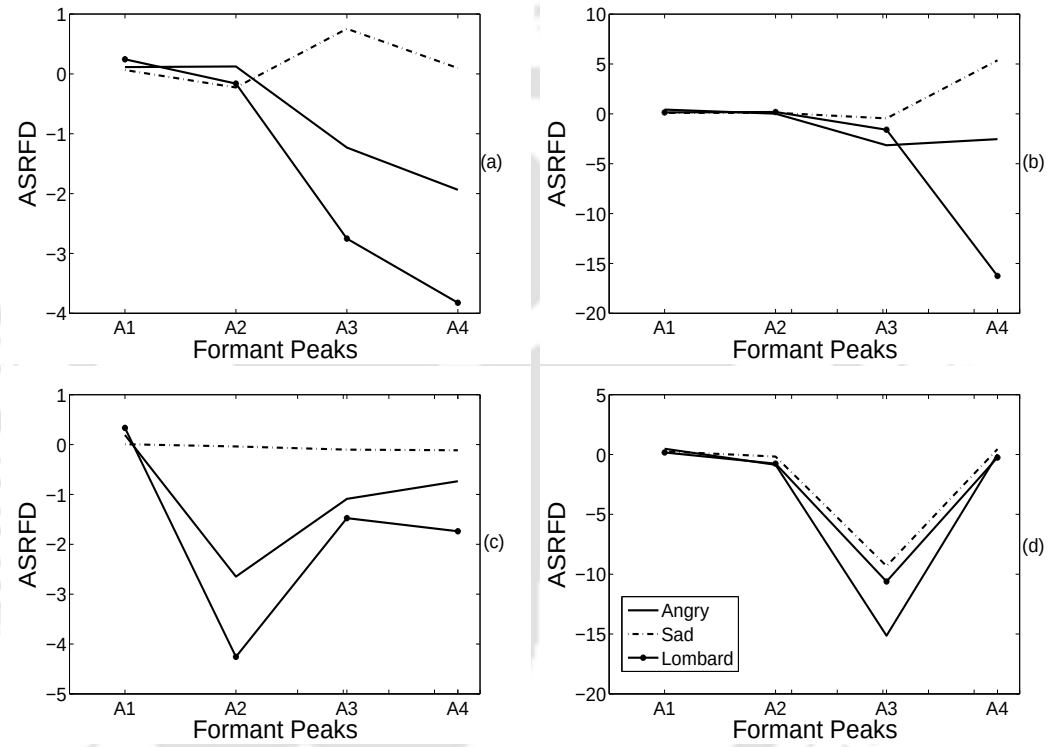

Figure 4.6: $ASRFD$ values of vowels under different stressed conditions (a) /a/ (b) /e/ (c) /i/ and (d) /u/

This study shows that 1^{st} formant peak is less effected by stress whereas the effect of stress is more on higher formant frequencies. Further, the variation of $ASRFD$ is distinct under different stressed conditions. This observation shows that the displacement of formant peaks from the 1^{st} formant peak, as described in previous subsection, carries systematic stress specific information.

4.3 Stress Classification Using RFD Feature

Section 4.1 shows that the gross spectral tilt has ability to characterize the stressed speech. Similarly, Section 4.2 shows that the RFD values contain stress specific information. This section investigates the stress information in gross spectral tilt and RFD values by considering them as a feature for stress classification. This feature is formed by concatenating three RFD values and one gross spectral tilt and the combined feature is termed as RFD feature. RFD feature contains slope and formant peak information of the spectrum. On the other hand, conventional $MFCC$ feature contains spectral energy information. Both information are estimated from the spectrum. In order to investigate the stress information in RFD feature, the performance of this feature is compared with the performance of $MFCC$ feature. Thirteen dimension mel frequency cepstral coefficient ($MFCC$) feature is extracted from the speech.

For evaluation of performance of stress classification, speech is divided into frames of length 160 samples (20 msec) and frame shift of 80 samples (10 msec). Tenth LP coefficients are calculated for each frame and the LP spectrum is computed. The displacements of 2^{nd}, 3^{rd} and 4^{th} formant peaks from 1^{st} formant peak, as described in Section 4.2, are computed for each frame. The LP spectrum of order one gives the overall slope of the spectrum which is approximately same as the slope of the spectrum computed by regression line method [2]. First order LP coefficient is concatenated with three RFD values estimated from LP spectrum. Hence, each frame is represented by four dimensional feature vector and it is termed as LP derived RFD feature (LP-RFD). Similarly, the RFD values from cepstral smoothed log spectra, as described in Section 4.2, are computed for each frame of the speech signal. The first order LP coefficient is concatenated with RFD values estimated from cepstrally smoothed spectrum and the combined feature is termed as log magnitude derived RFD (Log-RFD) feature. Approximately 4000 utterances (15,0000 frames approximately) taken from all speakers from 1^{st} session of database are considered for training and approximately 4000 utterances from 2^{nd} session of database are used for testing. The classifier is developed using continuous density HMM. Left to right (Bakis model) with no skipping in transition probability is considered.

Table 4.3: Performance of stress classification (%) for four class problem

	Neutral	Angry	Sad	Lombard	Avg. Perform
Listening test	51.29	73.86	88.25	31.39	61.20
LP-RFD	45.92	42.11	73.86	45.12	51.67
Log-RFD	48.53	40.95	72.21	47.91	52.40
$MFCC$	56.77	38.53	64.07	53.24	53.15
Feature level combination					
($RFD(LP)$-$MFCC$)(17dim)	59.45	47.51	66.74	53.02	56.68
($RFD(log)$-$MFCC$)(17dim)	60.00	50.00	66.59	53.95	57.63
Score level combination					
$RFD(LP)$-$MFCC$	56.33	48.68	67.25	55.35	56.90
$RFD(log)$-$MFCC$	57.56	47.29	64.55	57.67	56.67
Rank level combination					
$RFD(LP)$-$MFCC$	41.63	63.38	76.57	47.91	57.37
$RFD(log)$-$MFCC$	43.07	59.73	76.37	58.60	59.53

During the testing, Viterbi optimization criterion is used. In this work, 16 states are considered to model the stress class. Each state has four mixture components. The evaluation of this feature has been done as a four class problem. In this four class problem, the classifier has to decide the stress class of a given utterance among four stress classes, namely, neutral, angry, sad and Lombard. The performances of stress classification using $MFCC$ and RFD features are tabulated in Table 4.3. It is observed that the performances of neutral and Lombard speech using $MFCC$ feature are 56.77% and 53.24%, respectively. The performances of LP-RFD feature is reduced to 45.92% for neutral speech and 45.12% for Lombard speech. The performances of LP-RFD feature under angry and sad conditions are better than the $MFCC$ feature. Under angry and sad conditions, the performances of LP-RFD feature are 42.11% and 73.86%, respectively, which are 4.42% and 9.79%, respectively higher than the $MFCC$ feature. Similar results are observed for Log-RFD feature. The performances of Log-RFD feature under neutral and Lombard conditions are less by 8.24% and 5.24%, respectively than the $MFCC$ feature. On the other hand, the performances of Log-RFD feature under angry and sad conditions are more by 2.42% and 8.14%, respectively than that of $MFCC$ feature. Further, the average performance of Log-RFD feature is observed to be higher than that of LP-RFD

feature. This result infers that, the formant peak information of speech is more consistent in the cesptrally smoothed feature than LP based feature. This result also infers that the performance of four dimensional RFD feature is approximately similar to the performance of 13 dimensional $MFCC$ feature. This study shows that both the RFD feature and the $MFCC$ feature have approximately same stress discrimination capability. $MFCC$ feature captures spectral energies of the speech signal where formant peaks information are normalized. On the other hand, RFD feature captures formant peak information. Hence, the information present in $MFCC$ and RFD features are different and they may be capturing different aspects of stress information. The combination of these features may provide better improvement in the classification rate.

4.3.1 Feature Level Combination

As described in the previous section, the information present in the $MFCC$ and RFD features are different and they may reflect different aspects of stress specific information. Thus, the combination of these features in feature level may give better classification rate. In this study, 13-dimensional $MFCC$ feature is concatenated with 4-dimensional LP-RFD feature and Log-RFD feature, respectively. The 17-dimensional $RFD - MFCC$ combined feature is used to model stress. During the recognition, these 17-dimensional features are also used for stress classification. The performances of LP and log derived RFD features combined with $MFCC$ feature, are shown in Table 4.3. The performance of $RFD(LP)$-$MFCC$ feature is higher than their individual performances for all the stressed conditions except sad speech. Similar observation is found for $RFD(Log)$-$MFCC$ feature. The dimension of $MFCC$ feature is higher than that of RFD feature and the dynamic ranges of RFD features are significantly lower than that of $MFCC$ feature. Therefore, the contribution of $MFCC$ feature in recognition performance is more than that of RFD feature. Due to this, the performances of combined feature are more only for those stress conditions where the recognition performance of $MFCC$ feature is higher than the recognition performance of RFD feature. The average performances of $RFD(LP) - MFCC$ and $RFD(Log) - MFCC$ features are 56.68% and 57.63%, respectively

which are 3.53% and 4.48% higher than baseline performance by the $MFCC$ feature.

4.3.2 Score Level Combination

As described in subsection 4.3.1, the dynamic range of RFD feature is lower than that of $MFCC$ feature. The classification level combination may improve the performance of the combined feature. Two classification level combinations namely, score level and rank level are used in this work. In case of score level combination, the measurement scores of features are combined. The stress class with highest combined score is identified as the stress in the testing utterance. This study considers log likelihood value as a measurement score. The log likelihood score of each stress class for LP-RFD and Log-RFD features are combined with log likelihood score of each stress class for $MFCC$ feature. The classification performance of $RFD(LP) - MFCC$ and $RFD(Log) - MFCC$ features are shown in Table 4.3. Results show that the performances of $RFD(LP)$-$MFCC$ and $RFD(Log)$-$MFCC$ features are higher under neutral, angry and Lombard conditions as compared to their individual feature performances. For sad speech, the performances of these features are higher than those of $MFCC$ feature and they are less than those of RFD features. The average performance of combined $RFD(LP)$-$MFCC$ and $RFD(Log)$-$MFCC$ features are 56.90% and 56.67%, respectively, which are 3.75% and 3.52% higher than the baseline $MFCC$ feature.

4.3.3 Rank Level Combination

In case of rank level combination, a rank is given to the stress class according to its log-likelihood score. Among, the four stress classes, the highest rank is given to that stress class, which has highest log likelihood score. Hence, the ranks for four stress classes for RFD and $MFCC$ features are decided according to their log likelihood scores. Thus, there are two ranks for each stress, one using RFD feature and the other using $MFCC$ feature. The two ranks are summed to get the combined rank. The stress with highest combined rank is identified as the stress in the testing data. The performances of $RFD(LP)$-$MFCC$ and $RFD(Log)$-$MFCC$ features are shown in Table 4.3. The average performances of $RFD(LP)$-$MFCC$ and $RFD(Log)$-

$MFCC$ features are 57.37% and 59.53%, respectively. From this table, it can be seen that the performance of $RFD(Log)$-$MFCC$ feature under angry, sad and Lombard conditions are higher than their individual performances. On the other hand, the performances of $RFD(LP)$-$MFCC$ feature under angry and sad conditions are significantly improved, whereas the performance of $RFD(LP)$-$MFCC$ feature under Lombard condition is similar to the performance of LP-RFD feature. However, the performances of these features under neutral condition degrades from their individual feature performances. Although, the degradation is observed in the neutral speech performance, the average performances of $RFD(LP)$-$MFCC$ and $RFD(Log)$-$MFCC$ features improve by 4.22% and 6.38%, respectively from the performance of $MFCC$ feature.

In this section, the RFD feature is combined with the $MFCC$ feature at the feature level, at the score level and at the rank level. The stress classification performances of all these combination techniques show improvement from their individual feature performances. From these studies, it can be concluded that the RFD feature contains some additional information of stress, which is not present in the $MFCC$ feature.

4.4 Stress Dependent Speech Recognition

This section presents stress dependent speech recognition, as described in Subsection 1.2.2.2, to show that the stress class information obtained from proposed stress classifier provides improvement in the performance of the speech recognition. The block diagram of the stress dependent speech recognition is shown in Figure 4.7 [3]. In this technique, word model is trained for each stress during the training, therefore, the speech recognizer is developed for each stress class. Figure shows that the stress dependent speech recognizer is combined with the stress classifier. During the testing, the stress classifier identifies the stress class of the speech signal and the feature of the speech signal is tested using only the trained word models of the corresponding stress class. For each stress class, variation among the features is only due to the different words and hence an improved performance is expected during speech recognition.

In Chapter 3, the recognition performances of $MFCC$ and $LPCC$ features were evaluated

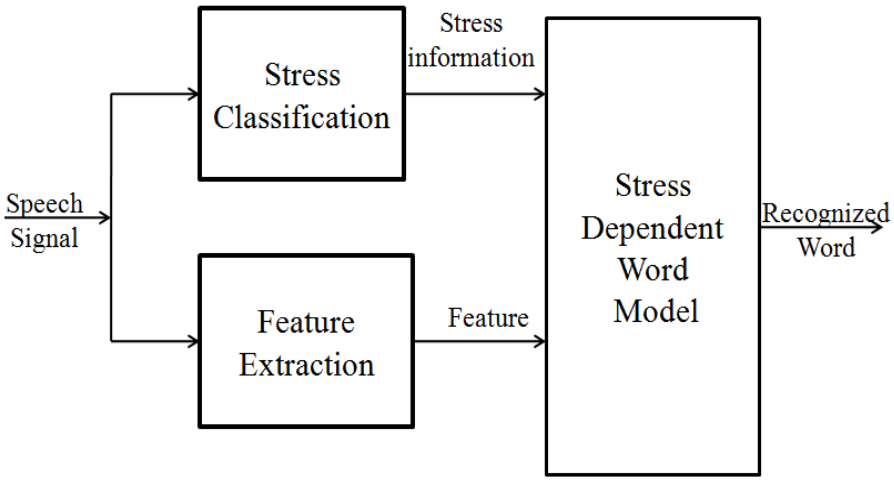

Figure 4.7: Block diagram of stressed speech recognition

and it was observed that the performance of speech recognition degrades under the stressed condition from the neutral speech.This study also revealed that the recognition performance of $MFCC$ feature is better than $LPCC$ feature under stressed condition. Therefore, in the present study, $MFCC$ feature is considered for stressed speech recognition. During the training, $MFCC$ feature is extracted from each word for different stress classes and each word model is trained for each stress class. During the testing, speech utterance is passed through the proposed stress classifier to identify the stress class, and speech recognition is carried out within that stress class. In this study, $MFCC$, $LP\text{-}RFD$ and $Log\text{-}RFD$ features are used for stress classification. Also, the rank level combination of $MFCC$ and RFD, as described in Subsection 4.3.3, are also considered for classification of stress class. The performances of stressed speech recognition using stress classification are compared with neutral trained speech recognizer.

The performances of $MFCC$ feature based speech recognition using different stress classification techniques are tabulated in Table 4.4. The average performance of speech recognizer for $LP\text{-}RFD$ and $Log\text{-}RFD$ features based stress classifier are 67.71% and 68.87%, respec-

Table 4.4: Performance of stress dependent speech recognition

	Neutral	Angry	Sad	Lombard	Avg. Perform
Neutral Trained	81.10	42.58	59.89	55.69	59.82
Feature for stress classification					
$MFCC$	70.32	67.63	63.44	73.36	68.69
$LP-RFD$	69.56	64.19	61.67	72.43	66.96
$Log-RFD$	69.14	64.14	63.44	72.91	67.40
$RDF(LP)-MFCC$	73.12	69.25	64.49	74.53	70.35
$RFD(log)-MFCC$	72.62	69.72	63.99	75.19	70.38

tively, which is 7.42% and 8.58% higher than that of neutral trained speech recognition. The performance of speech recognition for $MFCC$ based stress classifier is 68.69%, which is higher than that of RFD based stress classifier. The speech recognition performance depends on the classification rate of stress classifier. It is observed from Subsection 4.3, the classification rate of $MFCC$ feature is more than that of RFD feature, therefore the speech recognition using $MFCC$ based stress classifier gives better performance than that of RFD based stress classifier. From this table, it is observed that the performance of individual stress class, except for neutral speech, improves from that of neutral trained model. In neutral trained system, neutral speech is always tested with neutral speech. However, due to misidentification of stress classifier, neutral speech is tested with other stress trained speech recognizer. Hence, reduction in the recognition performance is observed in the neutral speech. In case of rank level combination, the average performances of speech recognition is 70.35% and 70.38% for $RFD(LP)$-$MFCC$ and $RFD(Log)$-$MFCC$ features, respectively, which are 10.53% and 10.56% higher than that of neutral trained speech recognition. As described in Subsection 4.3.3, the RFD feature contains some additional information of stress, which is not present in the $MFCC$ feature, which may contribute to the improvement of classification rate of combined feature. Therefore, the speech recognition performance using combined features gives better recognition performance as compared to their individual recognition performance.

4.5 Summary

This chapter demonstrates the effect of stress on the slope of the spectrum. Systematic variations of spectral tilt are observed under different stressed conditions. Studies show that the effect of stress are more pronounced at formant peaks. The gross slope of the spectrum can not capture this information effectively. Therefore, local spectral tilts are proposed in terms of relative displacement between formant peaks. Under the stressed condition, the RFD values show significant variation and the effect of stress is more for higher formant region. The RFD values are concatenated with gross spectral tilt and they together are used for classification of stressed speech. The performance of stress classification of RFD features, derived from LP spectrum and cesptrally smoothed spectrum, show that the performance of RFD feature is slightly less than traditional $MFCC$ feature, which shows that RFD feature has approximately the same discrimination capability for stress as $MFCC$ feature. Further, the performance of cepstrally smoothed log spectra derived RFD is higher than LP derived RFD feature. RFD features are combined with $MFCC$ feature at the feature level, at the score level and at the rank level. The results show that the combined features show marked improvement in performance as compared to their individual feature performances. The proposed stress classification techniques are employed to the speech recognition application improved the performance of recognition system.

5

Evaluation of Subband Energy

Contents

The effect of stress on slope and formant peaks of the spectrum is investigated in Chapter 4. The variation in slope reflects the migration of spectral energy from the lower frequency to the higher frequency. It was observed that the effect of stress is more on higher formants than on lower formants due to the migration of spectral energy and the variation in vocal loudness. The migration of spectral energy effects not only the features of the spectrum, but also the outputs of filterbanks which are placed in this spectrum [1] [98]. Stanton et al. [85] observed the trend of the energy migration in frequency domain for angry and Lombard speech. They observed that the additional spectral energy moves from low to mid bands frequency range which is sensitive to the human auditory system. Hansen and Womack [10] observed the migration of spectral energy in mel subbands. They observed maximum recognition performance near 2^{nd} formant region rather than 1^{st} formant region under angry, loud and Lombard conditions. To improve the recognition performances under these conditions, mel-filters were modified in such a way that the linear spaced filters occur near the 2^{nd} formant region. Thus, the modified mel filters gave more emphasis to the 2^{nd} formant region. Sarikaya and Gowdy [61] analyzed the effect of stress on subbands of the speech. Autocorrelations of energy of subbands were estimated to measure frame to frame correlation of energy and to model the relative change in subband energy due to stress. These studies show that the subbands get effected when speech is produced under stressed condition.

This chapter presents an analysis of effect of stress on subbands of speech. Effect of stress on individual subband is studied under different stressed conditions. Three statistical characteristics namely, means, variances and divergence of subbands are analyzed under different stressed conditions. Dynamics of subbands are also investigated to study the variation of energy across the subbands. Backward differences of subband energy is used to quantify the dynamics of subbands. Backward differences of subband energy based stress compensation techniques is proposed to investigate the robustness of this proposed analysis towards the stress. Four stress compensation techniques, based on statistical characteristics and dynamics of subbands, are proposed. The objective of these compensation techniques is to remove the effect of stress from subbands and the energy of subbands are assumed to be close to that of neutral speech. This

assumption is verified using stressed speech recognition.

The statistical analyses of the subbands under different stressed conditions are described in Section 5.1. Techniques to compensate the effect of stress from the subbands are discussed in Section 5.2. The effectiveness of these compensation techniques are evaluated in Section 5.3. The performances of stress compensation based speech recognition are evaluated in Section 5.4. Section 5.5 describes the summary of the chapter.

5.1 Analysis of Subband Energy for Different Stressed Conditions

As described in previous chapter, the migration of spectral energy depends on the glottal vibration. The migration of spectral energy for different stressed conditions is not the same, therefore, the subband energy (SBE) would be different for different stressed conditions. This section analyzes the effect of stress on subbands under different stressed conditions.

The probability density function (PDF) of subbands of speech is estimated under different stressed conditions to study the effect of different stressed conditions in statistical characteristics of SBE. In order to estimate the PDF of the subbands, the SBE are computed from 119 words vocabulary dataset under neutral, angry, sad and Lombard speech from 1^{st} session of database. The speech signal is segmented into a number of frames with length of 20 msec and frame shift of 10 msec. Hamming window is applied on these frames to reduce the effect of discontinuities at the end of the frames. The discrete Fourier transform (DFT) of each windowed frame is computed. The DFT spectrum is then multiplied with nonlinear mel-filter and the logarithmic magnitude of the filter output is termed as a *Subband Energies* of the speech signal. Approximately 50,000 frames per stress class are considered and 24 subbands are estimated from each frame. The maximum and minimum of each subband are computed and the difference between them are divided equally to find the number of bins, and the PDF of a subband is estimated by histogram method. In the subband energy computation, speech signals are considered for all the words from all the speakers present in 1^{st} session of database to smooth out the local effects, such as word and speaker, from the PDF of subband energy. Therefore,

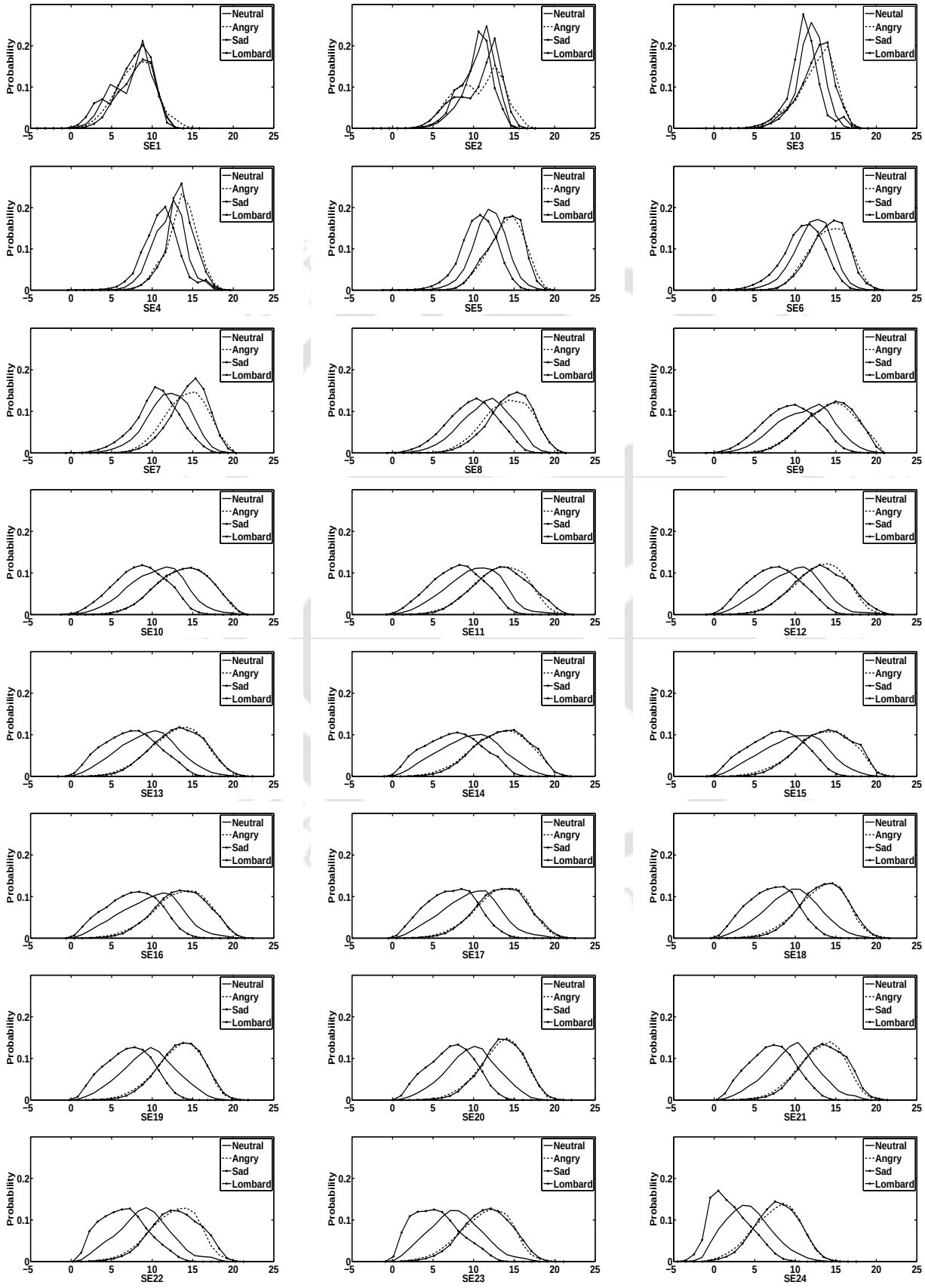

Figure 5.1: Probability density function of subband energy of speech under different stressed conditions

the PDF only represents the stress information present in this subband. Figure **??** shows the $PDFs$ of subbands under four stressed conditions namely, neutral, angry, sad and Lombard. Figure shows that the statistical characteristics of SBE of stressed speech significantly vary from the neutral speech. For instance, the means of SBE of stressed speech are shifted from the neutral speech and the $PDFs$ of SBE of speech under stressed and neutral conditions are significantly separated from each other. These variations in subbands under stressed conditions infer that the subband contains stress specific information, therefore, a detailed analysis of these statistical characteristics of SBE may provide useful information of stress.

5.1.1 Analysis of Statistical Characteristics of Subband Energy for Stressed Speech

Based on previous observations, in this section, three statistical parameters namely, mean, variance and divergence of subbands are analyzed for different stressed conditions. For a given stressed condition, the mean energy of a particular subband is the average energy of 50,000 frames in that particular subband. Similarly, the mean energy values of 24 subbands are computed for different stressed conditions and they are shown in Figure 5.2. For angry and

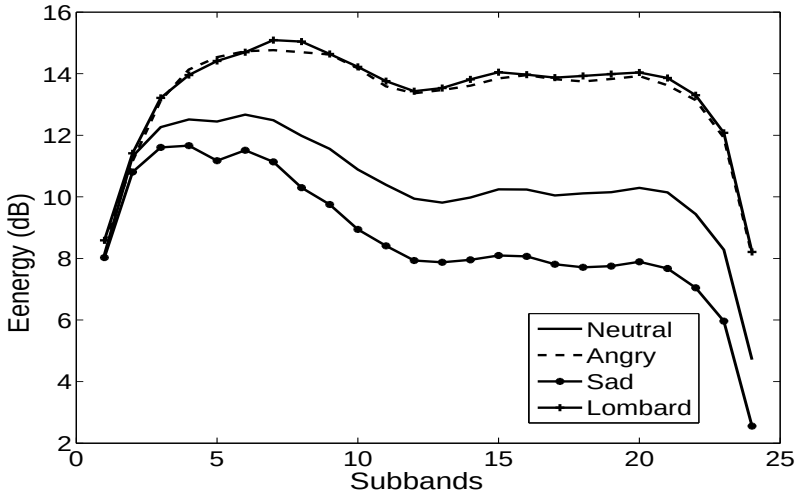

Figure 5.2: Mean of subband energy under different stressed conditions

Lombard speech, the vocal loudness and the rate of glottal vibration are higher than those for

neutral speech. As a result, more spectral energy is shifted to the higher subbands. Due to this migration of energy, the mean energy values of higher subbands are higher for angry and Lombard speech. On the other hand, the vocal loudness and the rate of glottal vibration for sad speech are comparatively lower than those for the neutral speech, which reduces the spectral energy of higher subbands. As a result, the mean energy values of higher subbands of sad speech are lower than that of neutral speech. The migration of spectral energy and the vocal loudness effect not only the mean of the individual subband under stressed condition, but also the slope of the SBE. Here, slope of the subbands is defined as the difference between mean energy of 24^{th} subband and mean energy of 1^{st} subband. From Figure 5.2, it can be observed that the slope of SBE is higher for angry and Lombard speech, and lower for sad speech as compared to the neutral speech.

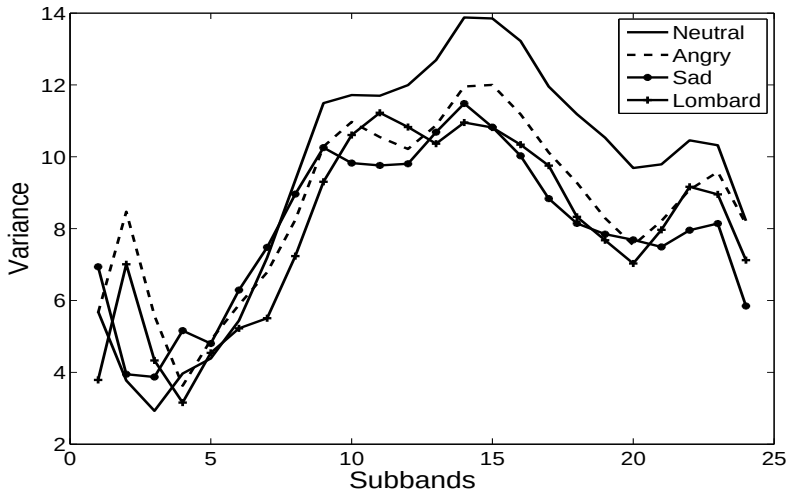

Figure 5.3: Variance of subband energy under different stressed conditions

The variance values of subband energy are computed from approximately 50,000 frames of a given stressed condition. The variances of SBE under different stressed conditions are shown in Figure 5.3. Figure shows that the variance values of higher subbands are reduced under stressed conditions as compared to the neutral condition. Further, the variation in variance values of the subbands for stressed speech are not consistent as observed from the mean energy values of subbands.

Figure **??** shows that SBE distribution under stressed condition is deviated significantly from the SBE distribution under neutral condition. The KL-divergence (KLD) between SBE distribution of stressed speech, $N(.;\mu_E^X,\sigma_E^X)$, from SBE distribution of neutral speech, $N(.;\mu_E^N,\sigma_E^N)$ is computed to determine how distinguishable the two distributions are [97]. Figure **??** shows that the distributions of SBE under stressed condition are normal (Gaussian) distributed. Therefore, the $\mathrm{D(N(.;\mu_E^N,\sigma_E^N)||N(.;\mu_E^X,\sigma_E^X))}$ for the distributions $N(.;\mu_E^N,\sigma_E^N)$ and $N(.;\mu_E^X,\sigma_E^X)$ is calculated by Eq. 5.1.

$$\mathrm{D(N(.;\mu_E^N,\sigma_E^N)||N(.;\mu_E^X,\sigma_E^X))} = \frac{1}{2}[\log(\frac{\sigma_E^{X2}(k)}{\sigma_E^{N2}(k)}) - 1 + \frac{\sigma_E^{N2}(k)}{\sigma_E^{X2}(k)} + \frac{(\mu_E^N(k)-\mu_E^X(k))^2}{\sigma_E^{X2}(k)}] \quad (5.1)$$

The divergences between distributions of SBE under different stressed conditions from the

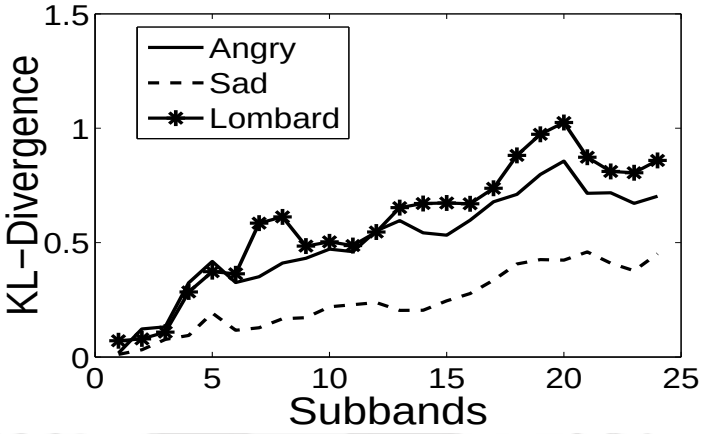

Figure 5.4: Divergence of SBE under different stressed conditions from neutral speech

neutral condition are shown in Figure 5.4. It is observed that, the deviations in distributions of higher subbands are comparatively higher than those of lower subbands. The analysis of mean and variance of subbands demonstrated that the difference of means of $f(x)$ and $g(x)$ i.e. $(\mu_E^N(k)-\mu_E^X(k))$ is high for higher subbands. Also, variances are not much affected due to stress. As a result, the divergences of stressed speech from the neutral speech increases at higher subbands. Further, for lower subbands, the divergences of angry and Lombard speech are comparatively higher than the divergence of the sad speech. For higher subbands, the

divergence of Lombard speech is higher than the divergences of angry and sad speech.

5.1.2 Analysis of Proposed Difference Energy of Subband for Stressed Speech

Figure 5.2 shows that the slope of mean energy values of subbands for stressed speech is different from that of neutral speech. In this section, the local slope of the subbands is proposed in terms of difference of energy between consecutive subbands. The local slopes may contain the dynamic information across the subbands. The dynamics of the subbands might give information of rate of migration of energy across frequency scale. Backward difference of subband energy is proposed to quantify the rate of migration of energy across frequency scale and it is termed as *Difference Energy* of Subband. Difference Energy of k^{th} subband is defined by equation

$$\Delta E^X(k) = E^X(k) - E^X(k-1) \qquad 2 \leq k \leq 24 \tag{5.2}$$

where,

$E^X(k)$: Energy of k^{th} subband of a given X stress class

$\Delta E^X(k)$: Difference energy of k^{th} subband of a given X stress class

For a given frame, the difference energy is estimated using backward differences of the 24 subband energy. Therefore, 23 difference energy values are obtained from each frame. The difference energy values are computed from 119 words vocabulary dataset under neutral, angry, sad and Lombard conditions from fifteen speakers. Approximately, 50000 frames per stress are considered to compute the mean difference energy of subbands. For a given stressed condition, the mean difference energy of a particular subband is the average difference energy of 50,000 frames in that particular subband. The mean difference energy values for different stressed conditions are shown in Figure 5.5. Figure shows that the lower subbands i.e. up to 12 subbands, the difference energy under stressed condition varies from the neutral condition. In mel-scale, first twelve subbands may correspond to the 1400 Hz frequency region, which may contain 1^{st} and 2^{nd} formant information. For first twelve subbands, the difference energy of stressed speech

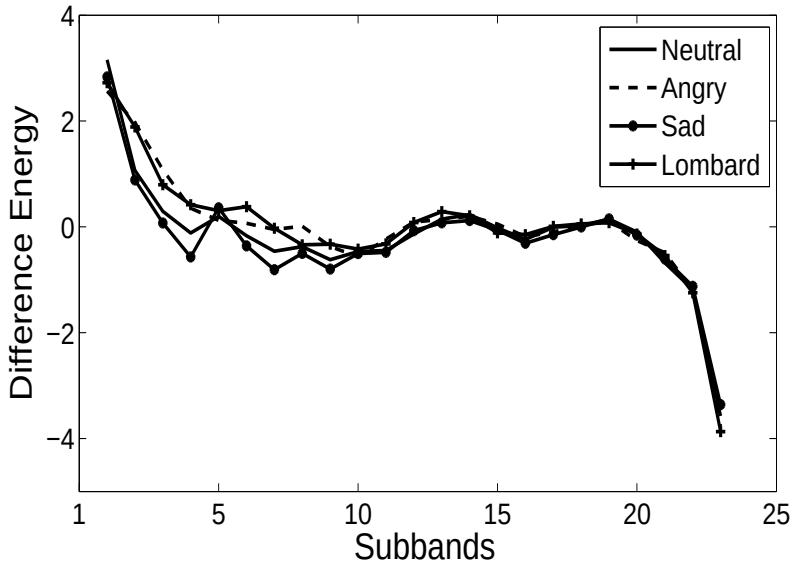

Figure 5.5: Difference energy of subbands under different stressed conditions

is compared with the difference energy of neutral speech. The difference energy is observed to be higher for angry and Lombard speech and lower for sad speech. This observation infers that the rate of migration of spectral energy from 1^{st} formant to 2^{nd} formant is higher for angry and Lombard speech than that for neutral speech. Also, the rate of migration of spectral energy from 1^{st} formant to 2^{nd} formant for sad speech is lower than that for neutral speech. This figure also shows that the difference energy of higher subbands are approximately same for all stressed conditions. This observation infers that the rate of migration of spectral energy is approximately constant for higher subbands under stressed condition. The difference energy of subband based analysis infers that the migration of energy takes place from 1^{st} formant to the 2^{nd} formant, which introduces modulation of energy at higher formants. Therefore, the energy of higher subbands under stressed condition are more deviated from neutral condition, as described in Section 5.1.

The PDF of difference energy is estimated by histogram method, as described in Section 5.1, for analysis of stress on the statistical characteristics of difference energy. Figure 5.6 shows the distributions of 4^{th}, 8^{th}, 12^{th}, 16^{th}, 20^{th} and 24^{th} subbands under different stressed conditions. The deviation in the distributions of lower subbands is observed under different

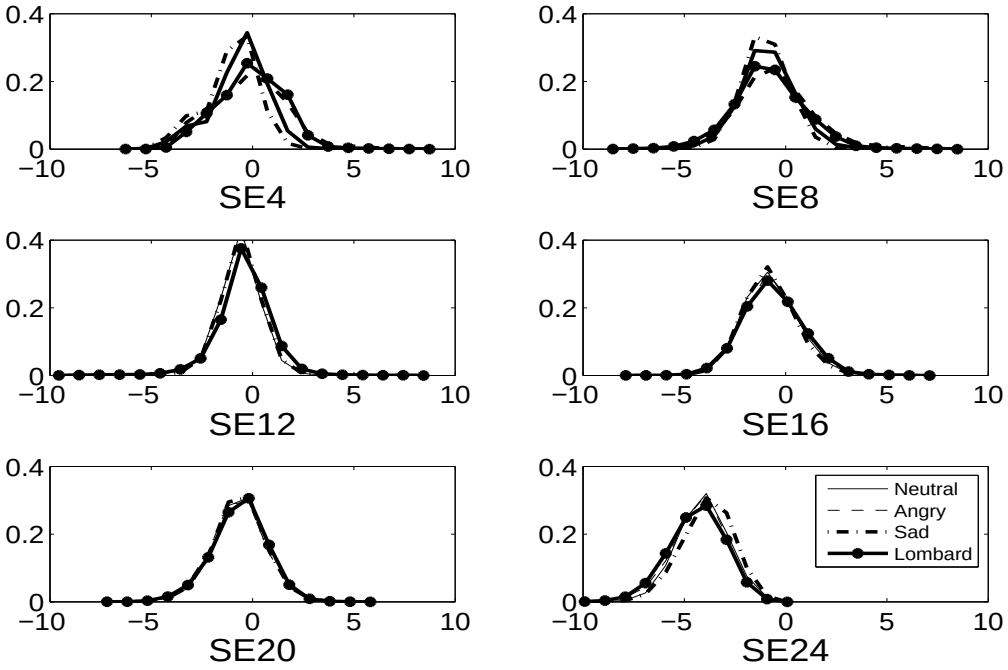

Figure 5.6: Probability density function of difference energy for different stressed conditions

stressed conditions. On the other hand, the deviation in the distributions of higher subbands is less. By comparing Figure 5.6 with Figure **??**, it can be observed that for stressed speech, the deviation in the distribution of difference energy is comparatively lesser than the deviation in the distribution of subband energy. This study infers that the SBE contains stress specific information, which is reflected from mean and divergence of the SBE. On the other hand, the difference energy of subbands is less effected by stress. Hence, difference energy of subbands may contain stress independent information.

5.2 Proposed Stress Compensation Techniques

Section 5.1 demonstrated the variations in statistical characteristics of subbands under stressed condition. These characteristics are the mean and the variance of subbands, the deviation of distribution of subbands and the dynamics of subbands. Some systematic pattern was observed in the variation of the characteristics of SBE under different stressed conditions, such as:

(i) Figure 5.2 shows that the mean energy of subband is more under angry and Lombard conditions than that under neutral conditions whereas, the mean energy of subband is lesser under sad speech than that under neutral conditions.

(ii) Figure 5.3 shows that the variance of higher subband energy is lesser under stressed conditions than that under neutral condition.

(iii) Figure 5.4 shows that the divergence values of higher subbands, under stressed condition, are higher as compared to the lower subbands.

(iv) Figure 5.5 shows that the dynamics of subband is invariant to the stressed condition.

In order to further investigate the effectiveness of these systematic patterns, stress compensation techniques are proposed. This section develops techniques to eliminate the effect of stress from the subbands. The objective of this compensation technique is to transform the energy of k^{th} subband, $E^X(k)$, of given stress X to modified energy of that subband, $\hat{E}^X(k)$, by using weighting factor $W^X(k)$ as given in Eq. 5.3.

$$\hat{E}^X = W^X E^X \tag{5.3}$$

The objective of this compensation technique is to transform SBE of stressed speech in such a way that the transformed SBE is close to SBE of neutral speech. In this study, four weighting factors are proposed to transform the SBE of stressed speech. These four weighting factors are weighted mean, weighted variance, normalized KLD and difference energy.

5.2.1 Weighted Mean

Figure 5.2 shows that the mean values of SBE vary under stressed conditions and the variations of mean values are observed to follow same systematic patterns. In order to normalize the effect of stress in the mean of SBE, weighting factor is proposed, which weights the means of SBE

under stressed condition. The weighting factor ($W_1X(k)$) of k^{th} subband is defined in Eq. 5.4.

$$W_1^X(k) = \frac{\frac{1}{M_1}\sum_{i=1}^{M_1} E_i^N(k)}{\frac{1}{M_2}\sum_{i=1}^{M_2} E_i^X(k)} \tag{5.4}$$

Here, M_1 and M_2 are number of frames for neutral speech and stressed speech, respectively and $E^X(k)$ is the energy of k^{th} subband for stress X. The weighting factors corresponding to angry, sad and Lombard speech are shown in Fig 5.7. The mean energy values of angry and Lombard speech are higher for higher subbands, therefore, the weighting factor of angry and Lombard speech decreases with increasing subbands to deemphasize the mean energy of higher subbands. On the other hand, the mean energy values for sad speech is lower for higher subbands, therefore, the weighting factor of sad speech increases with increasing subbands to emphasize the mean energy of higher subbands. The mean of SBE of a stress class is

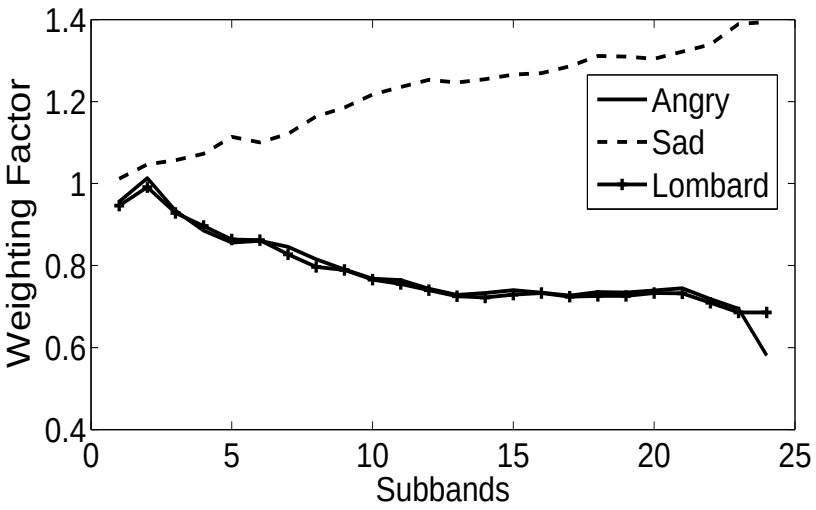

Figure 5.7: Weighting factor of weighted mean based compensation technique for different stress classes

the average of all other information such as speakers and words, therefore, it contains only stress information. The SBE of frames of an utterance may contain speech as well as stress information. The multiplication of the weighting factor of a stress class with the SBE of each frame of an utterance corresponding to that stress, as given in Eq. 5.3, may compensate the

effect of stress in the subband energy of the frames.

5.2.2 Weighted Variance

Figure 5.3 shows that the variances of higher subbands under stressed condition is lesser than that under neutral condition. On the other hand, the variances of lower subbands of the stressed speech are slightly higher than that under neutral speech. In order to normalize the effect of stress in the variance of SBE, weighted variance is proposed, which weights the variances of SBE under stressed condition. The weighting factor ($W_2X(k)$) of k^{th} subband is defined in Eq. 5.5.

$$W_2^X(k) = \frac{\sigma_E^N(k)}{\sigma_E^X(k)} \tag{5.5}$$

Here, $\sigma_E^N(k)$ and $\sigma_E^X(k)$ are the variance of subband energy of k^{th} subband for neutral condition and stressed condition, respectively. The weighting factors corresponding to angry, sad and Lombard speech are shown in Fig 5.8. The random variations in weights are observed for initial subbands (i.e. up to 10^{th} subbands). The weighting factor of a stress class is multiplied with the subband energy of frames of an utterance as given in Eq. 5.6.

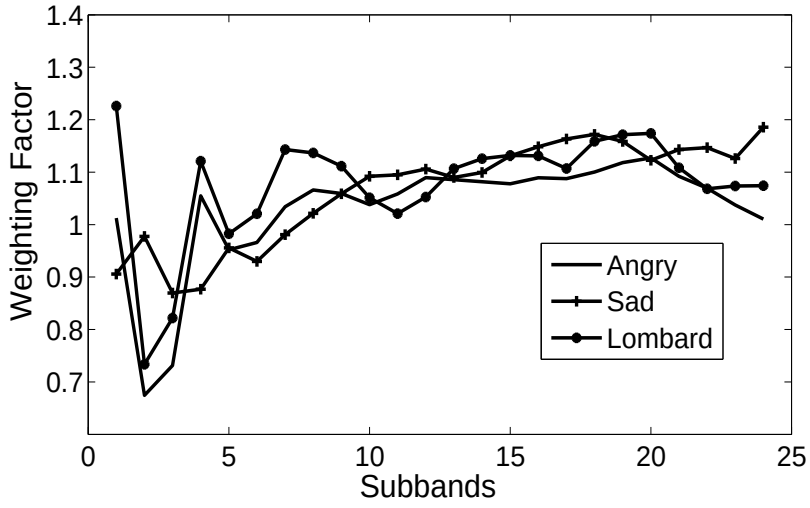

Figure 5.8: Weighting factor of weighted variance based compensation technique for different stress classes

$$\hat{E}_i^X(k) = W_2^X(k)(E_i^X(k) - \mu_E^X(k)) + \mu_E^X(k) \tag{5.6}$$

where,

$\sigma_E^X(k)$: Variance of subband energy of k^{th} subband for stress X

$W_2^X(k)$: Weighting factor for variance normalization of k^{th} subband for stress X

$\hat{E}_i^{X'}(k)$: Variance normalized SBE of i^{th} frame

5.2.3 Normalized Divergence

In this subsection, KLD information of subband energy is used to compensate the effect of stress in SBE of the speech signal. The weighting factor of k^{th} subband of a stress class is the normalized divergence value of that subband. The weighting factor of k^{th} subband is defined in Eq. 5.7.

$$W_3^X(k) = \begin{cases} (1 + \frac{D(k)}{max(D(k))}) & \text{for sad speech} \\ \frac{1}{(1+\frac{D(k)}{max(D(k))})} & \text{for angry and Lombard speech} \end{cases} \tag{5.7}$$

Here, $D(k)$ is the divergence of k^{th} subband from the stressed speech from the neutral speech. In Section 5.1.1, it was observed that the spectral energy of higher subbands is shifted upwards for angry and Lombard speech and downwards for sad speech. In order to deemphasize the energy of higher subband of angry and Lombard speech, the weighting factors of angry and Lombard speech are considered as inverse of their normalized KLD values. In case of sad speech, the normalized KLD values are used directly to boost the energy of the higher subband.

The normalized KLD based weighting factors for different stressed conditions are shown in Figure 5.9. Figure shows that for angry and Lombard speech, the weighting factor decreases as subband increases, whereas for sad speech, the weighting factor increases as subband increases. Also, the weighting factor is not uniform for all stress classes. Sudden variation in weights are observed for consecutive subbands. The multiplication of this weighting factor with the SBE of frames of stressed speech may introduce ripples in the cepstral domain. These ripples may interfere with the message information of the speech signal. In order to avoid sudden

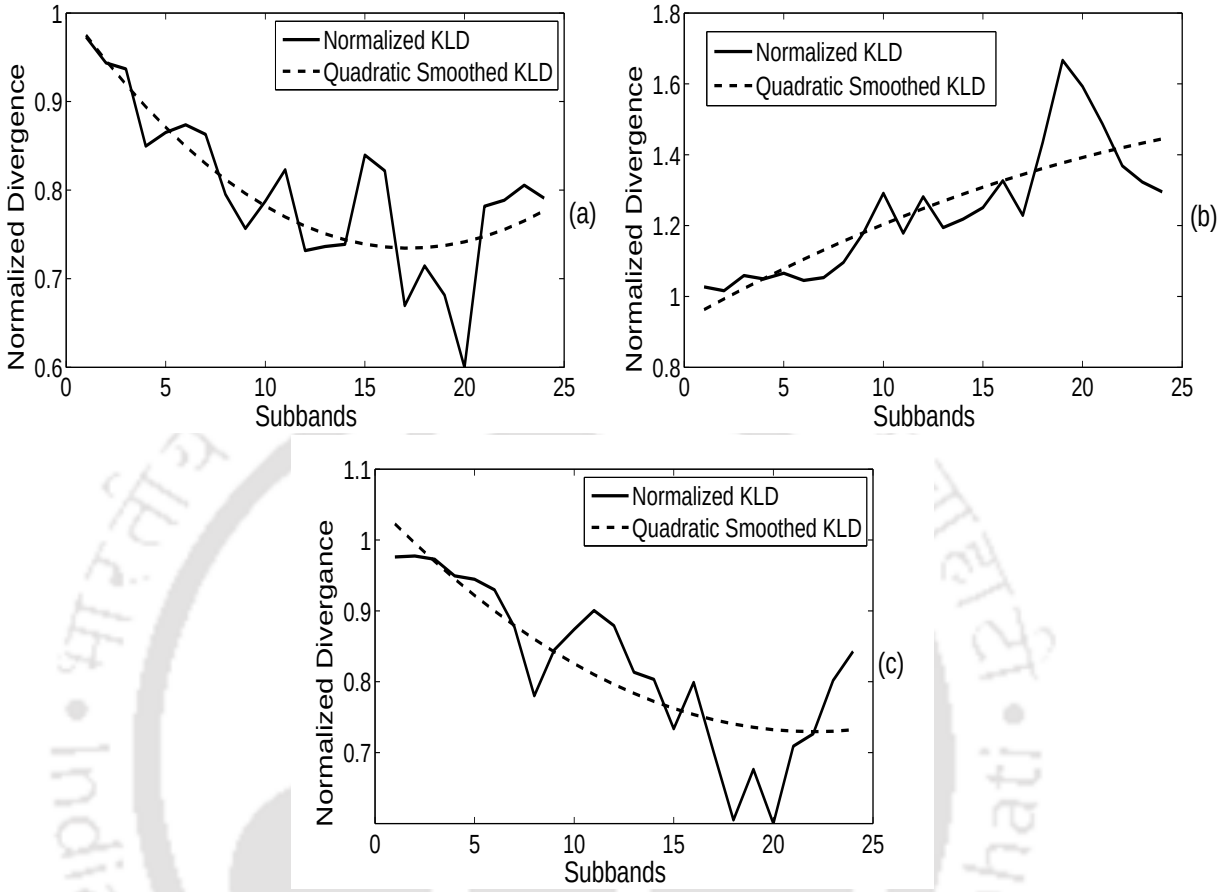

Figure 5.9: Weighting factor using normalized divergence and smoothed weighting factors for different stress classes, (a) angry (b) sad and (c) Lombard

variation in weighting factor, the smoothing of weighting factor by a quadratic function is done. The quadratic polynomial function is fitted to the weighting factor based on least square error method. The smoothed weighting factors of angry, sad and Lombard speech are given in Eq. 5.8.

$$W_4^X(k) = \begin{cases} 0.00092k^2 - .032k + 1 & \text{for angry speech} \\ -0.00041k^2 + 0.031k + 0.93 & \text{for sad speech} \\ 0.00066k^2 - 0.029k + 1.1 & \text{for Lombard speech} \end{cases} \tag{5.8}$$

The smoothed weighting factor of different stress classes are shown in Figure 5.9. The multiplication of smoothed weighting factor with the SBE of each frame of an utterance corresponding to that stress, as given in Eq. 5.3, is expected to compensate the effect of stress from

the SBE better than without smoothed weighting factor.

5.2.4 Difference Energy of Subband

This study investigates the stress insensitiveness of difference energy to compensate the effect of stress from the subbands. Figure 5.5 shows that the lower subbands i.e. up to 12 subbands, the difference energy under stressed conditions varies considerably from the difference energy under neutral condition, whereas for higher subbands (i.e. the last 10 subbands), the difference energy is approximately same for all stressed conditions. In order to compensate the variation in energy across 1^{st} and 2^{nd} formants, the difference energy knowledge is used as a mapping function for SBE of the stressed speech to SBE of the neutral speech. A recursive approach is proposed to map the SBE of the stressed speech to SBE of the neutral speech. Figure 5.2 shows that the mean energy of 1^{st} subband is approximately same for all stressed conditions. Therefore, the mean energy of 1^{st} subband of a stressed condition is considered for initialization of modified mean energy of 1^{st} subband. The initialization of recursive approach is given in Eq. 5.9. From difference energy analysis, as described in Section 5.1.2, it is observed that the energy of higher subbands are effected due to migration of spectral energy from 1^{st} formant to 2^{nd} formant regions. In order to compensate the effect of stress from these regions, the mean values of other subbands are recursively modified as given in Eq. 5.10. In this recursive approach, the difference energy of neutral speech is used to transform the mean energy of k^{th} subband of the stressed speech. The objective of the transformation is to make the transformed mean energy of subbands of stressed speech equal to the mean energy of subbands of neutral speech. The weighting factors of subbands are then estimated in Eq. 5.11.

(i) Initialization

$$\hat{\mu}_E^X(1) = \mu_E^X(1) \tag{5.9}$$

(ii) Recursive

$$\hat{\mu}_E^X(k) = \Delta E^N(k-1) + \hat{\mu}_E^X(k-1) \qquad 2 \leq k \leq 24 \tag{5.10}$$

(iii) Weighting factor

$$W_5^X(k) = \frac{\hat{\mu}_E^X(k)}{\mu_E^X(k)} \qquad 1 \leq k \leq 24 \tag{5.11}$$

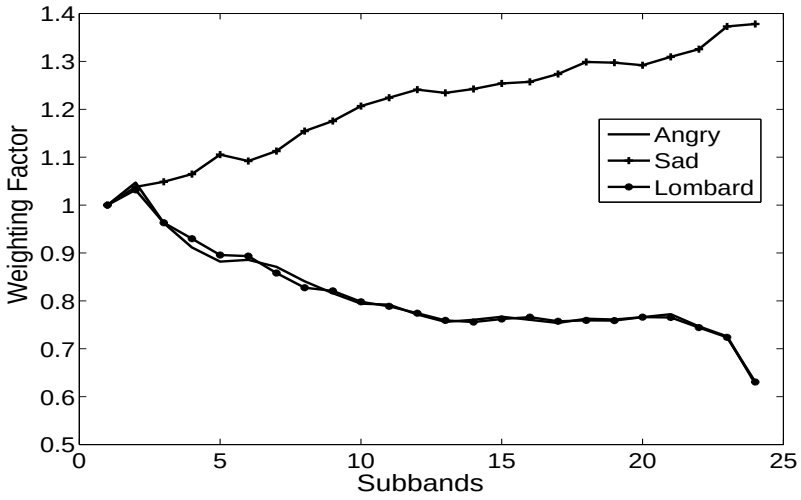

Figure 5.10: Weighting factor of difference energy based compensation technique for different stress classes

Here, $\hat{\mu}_E^X(k)$ is the compensated mean energy of k^{th} subband for stressed speech, and $\mu_E^X(k)$ is the mean energy of k^{th} subbands for stressed speech. $\Delta E^N(k)$ is the difference energy of k^{th} subband for neutral speech. The weighting factors of angry, sad and Lombard classes are shown in Fig 5.10. By comparing Figure 5.10 and Figure 5.7, it can be observed that the variation pattern of difference energy based weighting factor is approximately same as the variation pattern of weighted mean based weighting factor. This study infers that the mean of subband under stressed condition is approximately neutral after transformation of SBE using difference energy approach. Previous analyses (Section 5.1) show that the effect of stress is less in lower subbands, due to which, the difference energy based weighting factor gives weight approximately close to one for those subbands.

5.3 Evaluation of Proposed Stress Compensation Techniques

The proposed stress compensation techniques is expected to remove the effect of stress from subbands. In order to investigate this assumption, distance between compensated stressed feature and the neutral feature is computed. In this study, the distance between compensated stressed feature and neutral feature is measured in spectral domain. The effectiveness of these proposed stress compensation techniques are evaluated by performing comparison among all proposed stress compensation techniques.

5.3.1 Spectral Distance Measure

To evaluate the effectiveness of proposed compensation technique, the spectral distance between SBE of stressed speech and SBE of neutral speech is measured and expression of spectral distance is given in Eq. 5.12 [2]

$$d(\mu_E^N(k), \mu_E^X(k)) = |\mu_E^N(k) - \mu_E^X(k)|^2 \tag{5.12}$$

Here, $\mu_E^N(k)$ is the mean energy of k^{th} subbands for neutral speech, and $\mu_E^X(k)$ is the mean energy of k^{th} subbands for stressed speech. The subband energy is estimated from the speech under stressed condition (including neutral condition) of 2^{nd} session of database. Twenty four subband energy values are computed for approximately 4000 utterances (approximately 50,000 frames) per stressed condition. The average energy of these frames for k^{th} subband under neutral and stressed conditions are $\mu_E^N(k)$ and $\mu_E^X(k)$, respectively. The weighting factors of proposed compensation techniques, as described in Section 5.2, are computed for angry, sad and Lombard classes from 1^{st} session of database. The multiplication of proposed weighting factor (W_1^X) with the SBE of each of the frames corresponding to that stress, as given in Eq. 5.3, gives modified subband energy, $\hat{E}^X(k)$. The average energy of these frames for k^{th} subband is $\hat{\mu}_E^X(k)$. Similarly, $\hat{\mu}_E^X(k)$ is computed for each of the weighting factors that is W_2^X, W_4^X, and W_5^X. The spectral distance between $\mu_E^X(k)$ and $\mu_E^N(k)$ is computed using Eq. 5.12 and the spectral distances of angry, sad and Lombard speech are shown in Figure 5.11. Similarly,

the spectral distance between $\hat{\mu}_E^X(k)$,computed from each of the weighting factors, and $\mu_E^N(k)$ is computed using Eq. 5.13. The spectral distances of angry, sad and Lombard speech are also shown in Figure 5.11. Figure shows that the spectral distance is observed to be more at higher subbands under stressed condition when no weighting factor is used. The higher distances at higher subbands infers that the effect of stress is more at those subbands.

$$d(\mu_E^N(k), \hat{\mu}_E^X(k)) = |\mu_E^N(k) - \hat{\mu}_E^X(k)|^2 \tag{5.13}$$

Figure also shows that the modified subband energy, using weighted mean based compen-

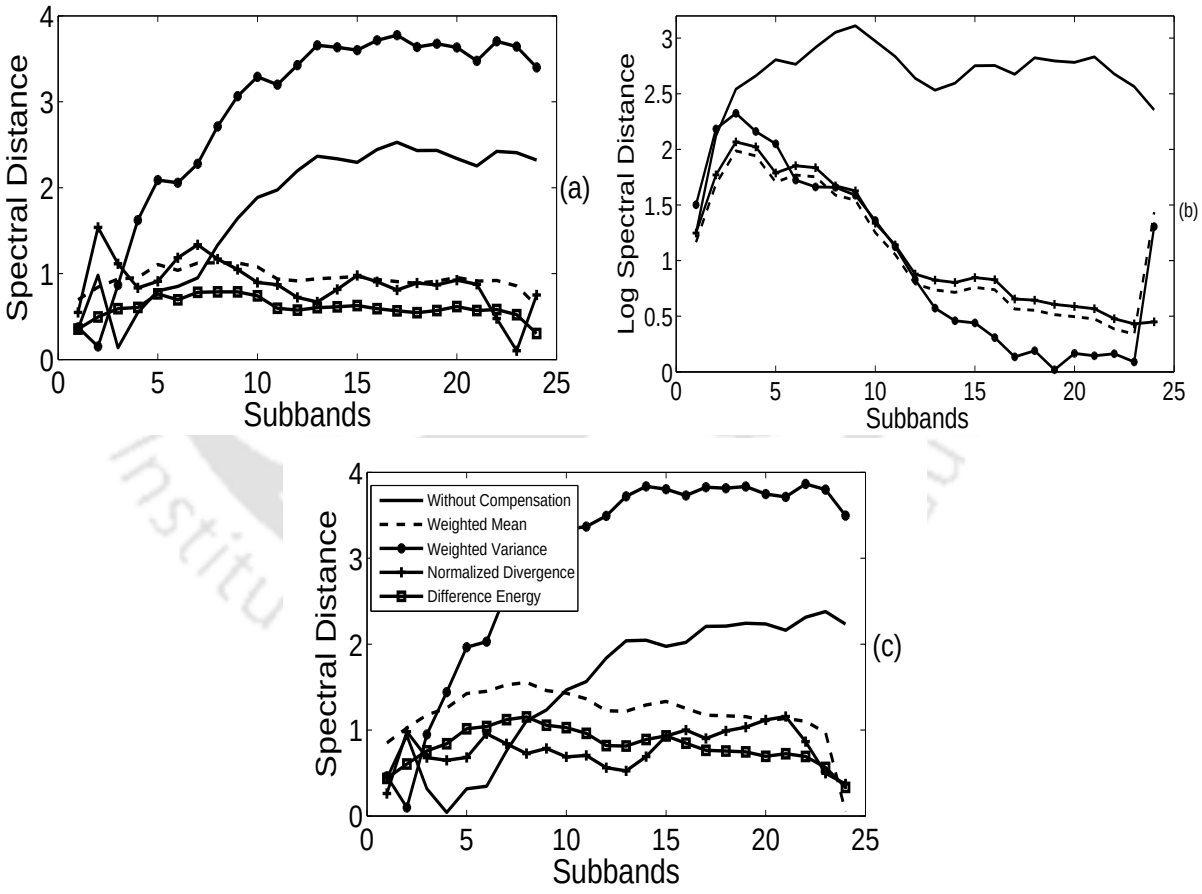

Figure 5.11: Spectral distances of different stressed speech from the neutral speech. (a) angry (b) sad and (c) Lombard speech

sation technique, gives approximately constant spectral distance across subbands for angry and Lombard speech. This observation infers that after compensation, the effect of stress on

SBE are removed effectively from higher subbands. However, the distances of lower subbands are comparatively higher than that of without compensated SBE of speech. Alternatively, the modified subband energy, using weighted variance based compensation technique, gives more spectral distance across subbands as compared to SBE without using any compensation technique. This observation infers that the weighted variance based compensation technique introduces distortion in the SBE after compensation. The modified subband energy, using normalized divergence based compensation technique, follows similar trend as weighted mean based compensation technique. For angry and sad speech, the distance of lower subbands of this technique are comparatively higher than that of weighted mean based compensation. It is also observed from the figure that difference energy based compensation technique follows similar trend as weighted mean based compensation technique, but at reduced distance level. This observation infers that the modified SBE using difference energy based compensation technique is closer to the SBE of neutral speech rather than using the weighted mean based compensation technique. From spectral distance based analysis, it can be concluded that the insensitiveness towards the stress of difference energy, provides more useful information to the compensator to remove stress information rather than other compensation technique.

5.4 Stressed Speech Recognition using Stress Compensation Techniques

In this section, the effectiveness of proposed compensation techniques to remove stress information from the subband is evaluated using speech recognition. The block diagram of proposed stress compensation technique based speech recognition is shown in Figure 5.12. In this figure, the subbands are computed in spectral energy reformation block. Based on statistical and dynamic analyses of subband, the spectral domain compensation techniques are proposed using weighting factors. The objective of computing weighting factor is to emphasize or deemphasize the energy of subband depending upon the level of stress on that subband. The modified energy values are considered for DCT computation to compute modified cepstral coefficients. The modified cepstral coefficients are tested by considering HMM based speech recognizer,

as described in Section 3.3. The performances of compensation techniques are evaluated by extracting the cepstral coefficients from the compensated SBE. The modified cepstral feature is compared with the $MFCC$ feature. As described in Section 3.3, the recognition performance of stressed speech using $MFCC$ as a feature is considered as baseline performance.

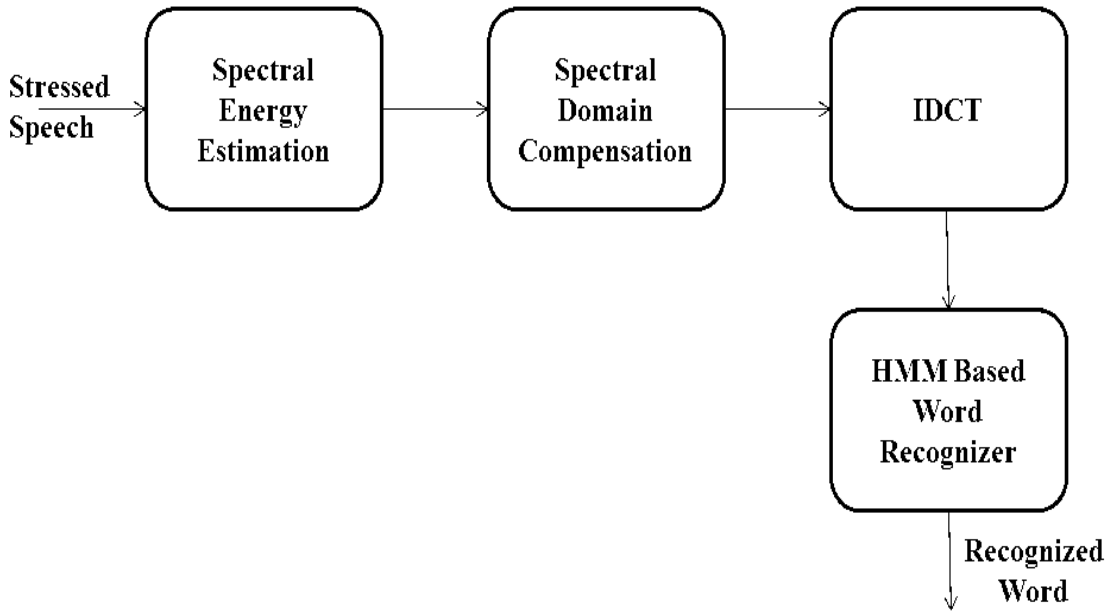

Figure 5.12: Block diagram of proposed compensation technique based stressed speech recognition

5.4.1 Ideal Stress Classifier

In this subsection, it is assumed that the ideal stress classifier is available to the recognizer during testing i.e. the prior knowledge of stress is given to the recognizer. The recognition performances of modified cepstral feature using proposed compensation techniques are tabulated in Table 5.1. The performances of cepstral feature, extracted using proposed weighted mean based compensation technique, under different stressed conditions are tabulated in the 3^{rd} row of Table 5.1. The recognition performances of angry and Lombard speech are observed to be 53.68% and 66.45%, respectively, which are 11.1% and 10.76% higher, respectively than

the baseline recognition performance. On the other hand, the recognition performance for sad speech is observed to be 58.34%, which is 1.55% less than the baseline performance. The average recognition performance of weighted mean based cepstral feature improves by 5.07% from the baseline performance.

Table 5.1: Performances of proposed stress compensation techniques based stressed speech recognition (in%)

	Neutral	Angry	Sad	Lombard	Avg. Perform
Baseline performance	81.13	42.58	59.89	55.69	59.82
Weighted mean	-	53.68	58.34	66.45	64.89
weighted variance	-	40.71	60.57	53.68	59.08
Normalized divergence	-	34.78	29.58	36.72	45.54
Smoothed weighting factor	-	53.25	53.76	64.66	63.19
difference energy weighting factor	-	53.85	59.24	67.05	65.31

Similar to the weighted mean based compensation technique, the recognition performances of weighted variance based cepstral feature under different stressed conditions are given in 4^{th} row of Table 5.1. The recognition performance of this feature degrades from the baseline performance. From Figure 5.3, it can be seen that the variance of stressed speech does not change for most of the subbands. Due to this, the feature may lose the speech specific information and hence, the performance degrades from the baseline recognizer. The normalized divergence based compensation technique is also evaluated for stressed speech recognition. The recognition performance of weighting factor, given in Eq. 5.7, based compensation technique, for different stressed conditions and their recognition performances are tabulated in 5^{th} row of Table 5.1. The recognition performance of this feature drastically degrades from the baseline feature. The recognition performance of smoothed weighting factor, given in Eq. 5.8, based compensation technique, for different stressed conditions are tabulated in 6^{th} row of Table 5.1. The smoothed weighting factor based compensation technique shows improvement in the recognition performance. The recognition performances of angry and Lombard speech are observed to be 53.25% and 64.66%, respectively, which are 10.67% and 8.97% higher, respectively than the baseline recognition performance. On the other hand, the recognition performance for sad speech is

observed to be 53.76%, which is 6.13% less than the baseline performance. The average performance of normalized divergence based cepstral feature improves 3.37% from the baseline performance. This result infers that due to the sudden variation in the weighting factor, as described in Subsection 5.2.3, the modified cepstral coefficients get distorted. As a result, the recognition performance degrades. On the other hand, after smoothing the weighting factors using quadratic polynomial, the discontinuity is smoothed out and recognition performance improves. From these results, it is also inferred that the subband energy contains dynamic information, which gets distorted when without smoothed normalized divergence based weighting factor is used. The recognition performance of modified cepstral feature, using difference energy based compensation technique, for different stressed conditions are given in 7^{th} row of Table 5.1. The performances of angry and Lombard speech are observed to be 53.85% and 67.05%, respectively, which are 11.27% and 11.36% higher, respectively than the baseline recognition performance. On the other hand, the the performance for sad speech is observed to be 59.24%, which is 0.65% less than the baseline performance. The average performance of difference energy based compensation technique based cepstral feature improves by 5.49% from the baseline performance.

From these results, it is observed that the performances of proposed compensation techniques improve from the baseline system. However, their performances still are not at par with the recognition performance of neutral speech. To understand the behavior of SBE after compensation, the spectral distance of SBE is considered, which is described in Section 5.3.1. The spectral distance of compensated SBE of stressed speech and SBE of neutral speech are shown in Figure 5.11. From this figure, it can be observed that the spectral distance is more at higher subbands for no compensation was done. The weighted mean based compensation technique gives approximately uniform spectral distance across subbands. After compensation, the effect of stress are reduced effectively at higher subbands. On the other hand, for all the stressed conditions, the spectral distances for lower subbands are comparatively higher than those without compensation. The spectral distortion of lower subbands may restrict further improvement in the recognition performance. The same observation is observed in normal-

ized divergence based compensation technique. The distance of lower subbands of normalized divergence based compensation are comparatively higher than the weighted mean based compensation. The performance of this technique is not at par with weighted mean based compensation technique. On the other hand, difference energy based compensation technique follows approximately the same trend as weighted mean based compensation technique. Therefore, difference energy based compensation technique gives comparatively higher performance than other compensation techniques.

5.4.2 Proposed Stress Classification Technique

Previous subsection evaluated the speech recognition using proposed stress compensation techniques. This compensation technique based speech recognition system assumes to have prior knowledge of stress class. However, the stress information is not known to the recognizer in a practical scenario. In this subsection, the proposed stress classification techniques, as discussed in Chapter 4, is considered as a pre-processor for speech recognition. In this study, the word model is developed using neutral speech and during testing, stress classifier identifies the stress class. The proposed weighting factor corresponding to the identified stress class is used to compensate the effect of stress from the subbands and modified subbands are used for cepstral computation. The modified cepstral feature are then used for recognition of speech.

In Section 4.3, it is observed that the performance of rank level combination of RFD feature, extracted from the LP and log smoothed spectra, with $MFCC$ feature gives better performance than their individual feature performances. Therefore, in this section, the performance of stress compensation techniques are evaluated only for rank level combination based stress classifier. From Subsection 5.4.1, it is observed that weighted mean, normalized divergence and difference energy based compensation techniques perform better than other compensation techniques. Hence, the performances of these techniques are evaluated using proposed stress classifier. The recognition performance of stress directed compensation based speech recognition is tabulated in Table 5.2. Table shows that the performances of weighted mean, normalized divergence and difference energy based compensation techniques are 64.89%, 63.19%, and 65.31%, respectively

Table 5.2: Performance of stress directed stress compensation based speech recognition

	Neutral	Angry	Sad	Lombard	Average performance
Neutral Trained	81.10	42.58	59.89	55.69	59.82
Ideal stress classification system					
Weighted mean	-	53.68	58.34	66.45	64.89
Normalized divergence	-	53.25	53.76	64.66	63.19
difference energy based weighting	-	53.85	59.24	67.05	65.31
$RFD(LPC)$-$MFCC$ feature for stress classification					
Weighted mean	75.03	52.10	58.61	64.30	62.51
Normalized divergence	74.81	49.50	57.77	62.37	61.11
difference energy based weighting	76.49	51.95	59.48	64.90	63.20
$RFD(Log)$-$MFCC$ feature for stress classification					
Weighting mean	75.23	52.11	58.43	65.05	62.63
Normalized divergence	74.94	49.46	57.07	63.07	61.14
difference energy based weighting	76.73	52.11	58.87	65.63	63.34

for ideal stress classification system. The performances of these techniques are reduced to 62.51%, 61.11% and 63.31%, respectively for $RFD(LP) - MFCC$ based stress classification. Similar observations are observed for $RFD(Log) - MFCC$ stress classification. Table 4.3 shows that the average stress classification rate of proposed rank level combination techniques, that is, $RFD(LP) - MFCC$ and $RFD(Log) - MFCC$ are 57.37% and 59.53%, respectively. This study shows that the rank level combination techniques, $RFD(LP) - MFCC$ and $RFD(Log) - MFCC$, fails 42.63% and 40.47% of the times, respectively to classify the stress class in the utterance. The multiplication of subband energy with weighting factor corresponding to the miss identified stress class produces incorrect compensation.

5.5 Summary

This chapter investigates that the effect of stress on subband energy of the speech signal. Statistical based analysis namely, mean, variance and divergence of SBE under different stressed conditions are considered. From these analyses, it is observed that the statistical characteristics such as mean and KL-divergence of SBE of the stressed speech vary from the neutral

speech. Alternatively, difference energy across subbands shows invariance nature for the stress. These knowledge have been further explored to compensate the effect of the stress and the modified feature is used for stressed speech recognition. The weighted mean based compensation technique provides good improvement in the performance. From the KL-divergence based study, it is found that the discontinuity in weighting factor introduces distortion in the cepstral features. This observation reveals that dynamic information across the subband is present, which gets distorted by the normalized weighting factor. In order to investigate the dynamic information of subbands, difference energy based compensation technique has been proposed for stressed speech recognition. The difference energy based compensation technique provides better recognition performance than the weighted mean based compensation technique. From the difference energy analysis, it is also observed that the difference energy is more robust for the stressed condition as compared to subband energy. The speech recognition performance using proposed rank level combination based stress classifiers shows that the performance of stress compensation depends on the classification rate of stress classifier.

6

Stress Analysis using Subspace Projection

Contents

It was observed in Chapter 5 that the migration of spectral energy effects subband energy of the stressed speech. The cepstral coefficients estimated from these subbands varied from cepstral coefficients of the neutral speech, due to which, the performance of speech recognition degraded. In the previous chapter, four stress compensation techniques are proposed to remove the stress from the subbands. However, all these techniques require additional knowledge of stress class. The focus of the present chapter is to develop a stress classifier which does not require explicit knowledge of stress class. Several stress compensation techniques have been proposed to remove the additional stress component present in the speech signal [11,12], [59]. In all these studies, the stress component was assumed as deterministic and additive at different levels of speech events, such as a broad phoneme [12] and at a word levels [11]. To eliminate this stress component, adaptive cepstral mean normalization (CMN) technique was proposed at the word and the broad phoneme levels [11, 12]. Afify et al. [59], assumed stress component as the additive random bias at the state level in continuous hidden Markov model ($CDHMM$) framework. In addition, the speech and the stress components were assumed statistically independent. Maximum likelihood state based additive bias model compensation technique was proposed to eliminate the additive random bias. From these studies, it can be observed that the separation of stress information from the speech signal at the feature level and the model level provided improvement in the speech recognition performance.

A subspace decomposition technique was proposed for separation of speech from the noisy speech signal [99]. This technique was proposed for the additive noise [99] [100], coloured noise [101] and speech enhancement for speech recognition [102]. In this technique, eigenvalue decomposition (EVD) and singular value decomposition (SVD) of the covariance matrices were used. Subspace projection technique was also proposed to increase the discrimination between the words [103]. In this study, subspace projection was achieved by divergence measure. These studies revealed that subspace projection technique has ability to separate the speech from the noisy speech [99] [101] and provide discrimination among words [103]. Similar to noisy speech signal, the stressed speech signal contains both speech and stress information. Literature showed that the performance of the speech recognition system can be improved if the stress information

is separated from the speech signal [11], [12], [59]. Several works have been reported for removal of stress from the speech signal at the feature level [11], [12] and the model level [59]. However, the effectiveness of subspace projection for separation of stress from the speech signal has not been analyzed. An analysis of stressed speech signal in this direction may help improve the performance of speech recognition system. In this work, subspace projection approach is proposed for analysis of stressed speech signal. Orthogonality relation is assumed to separate the speech and the stress information from the stressed speech signal.

The principle of subspace projection based approach is described in Section 6.1. The orthogonal relation between speech and stress information is verified in Section 6.2. The proposed subspace projection based speech recognition is described in Section 6.3. The proposed technique for speech recognition application is evaluated in Section 6.4. Section 6.5 describes the summary of the chapter.

6.1 Subspace Projection of Stressed Speech

This section presents the proposed subspace projection based approach to separate speech and stress information. This method assumes a linear model of neutral speech vector of dimension M where the neutral speech vector ($\mathbf{s}$) can be represented as

$$\mathbf{s} = \sum_{m=1}^{M} w_m \mathbf{v}_m \tag{6.1}$$

where, $w = (w_1, w_2, \ldots, w_M)$ are weights and $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_M]$ are M basis vectors [99]. Each basis is a K dimensional vector. The set of neutral speech vectors $\{\mathbf{s}\}$ can be assumed in a subspace of $\mathbf{R}^M$ spanned by the column of $\mathbf{V}$. This subspace is referred to as *Neutral Speech Subspace*. The vectors of this subspace are assumed to contain speech (semantic) information. The covariance matrix of the neutral speech vector $\mathbf{s}$ is given as

$$\mathbf{C}_s = E\{\mathbf{s}\mathbf{s}^T\} = \mathbf{V}\mathbf{C}_w\mathbf{V}^T \tag{6.2}$$

where, $\mathbf{C}_w$ denotes the covariance matrix of vector w. If the number of basis vectors is smaller than the dimension of the basis vector, the covariance matrix of speech vector, $\mathbf{C}_s$ contains

$K-M$ zero eigenvectors. Under stressed condition, the speech vector contains additional stress information. This vector can be termed as stressed speech vector. The stressed speech vector ($\mathbf{y}$) can be represented as a function (f) of speech ($\mathbf{s}$) and stress ($\mathbf{x}$) information. Mathematically, this can be written as $\mathbf{y} = f(\mathbf{s}, \mathbf{x})$. The set of stressed speech vectors $\{\mathbf{y}\}$ spans the stressed speech subspace. The covariance matrix of $\mathbf{y}$ can be considered as summation of covariance matrix of speech and covariance matrix of stress, if stress is assumed additive and orthogonal with the speech component in stressed speech vector [99]. It can be represented as

$$\mathbf{C}_y = E\{\mathbf{y}\mathbf{y}^T\} = \mathbf{C}_s + \mathbf{C}_x \tag{6.3}$$

where, $\mathbf{C}_x$ is the covariance matrix of stress component $\mathbf{x}$. The covariance matrix of stressed speech vector may also contain $K-M$ eigenvectors. These eigenvectors may contain stress information. Eigenvalue decomposition (EVD) can decompose the stress and speech components from $\mathbf{C}_y$ as given in

$$\mathbf{C}_y = \mathbf{U}\mathbf{\Lambda}_y\mathbf{U}^T \tag{6.4}$$

where, $\mathbf{U} = [\mathbf{u}_1, \ldots, \mathbf{u}_k]$ denotes the orthogonal matrix of eigenvectors of $\mathbf{C}_y$ and $\mathbf{\Lambda}_y = diag(\Lambda_y(1), \ldots, \Lambda_y(k))$ denotes the diagonal matrix of eigenvalues of $\mathbf{C}_y$. The subspace based methods are based on partitioning of the eigenvectors into a set belonging to the speech subspace, spanned by the columns of $\mathbf{U}$ and an orthogonal complement known as the stress subspace. This type of approach is normally used for noise filtration [99]. EVD can decompose the speech and stress subspaces more reliably when the covariance matrix ($\mathbf{C}_x$) of stress $\mathbf{x}$ is known. However, in the present case the stress information is not known, and therefore, the reliable decomposition of speech and stress subspace using this approach is not possible. In the present study, speech and stress components are assumed as orthogonal. To separate the stress and speech components from the stressed speech vector, subspace projection based approach is proposed [104]. In the present work, the geometrical properties of matrix-valued statistic is exploited to estimate speech and stress components from stressed speech vector. All vectors of neutral speech subspace of $\mathbf{R}^M$ are represented by a set of representative vectors termed as codevectors. The codevectors $\mathbf{A} = \{\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_N\}$ are the mean vectors. The codevectors of

neutral speech subspace are linear combinations of speech vectors, which are spanned by the column of $\mathbf{V}$. Therefore, these vectors lie in the subspace of $\mathbf{R}^N$.

The subspace projection based analysis is shown in Figure 6.1. Figure shows that the stressed speech vector,$\mathbf{y}$, deviates from the neutral speech subspace, $\{\mathbf{a}_n\}$. In this study, this deviation is assumed due to the stress present in the signal. Let, $\mathbf{P}_n$ be the projection matrix required to project a vector onto $\{\mathbf{a}_n\}$. $\mathbf{P}_n$ is given as

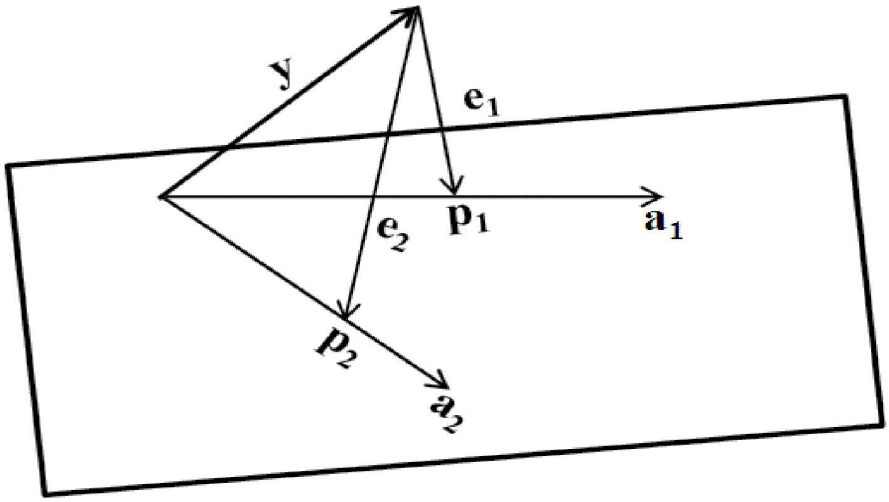

Figure 6.1: Subspace projection based analysis.

$$\mathbf{P}_n = \frac{\mathbf{a}_n \mathbf{a}_n^T}{\mathbf{a}_n^T \mathbf{a}_n} \qquad 1 \leq n \leq N \tag{6.5}$$

The multiplication of $\{\mathbf{P}_n\}$ with $\mathbf{y}$ produces projection of $\mathbf{y}$ onto $\{\mathbf{a}_n\}$, which can be represented as $\mathbf{p} = \{\mathbf{P}_n\mathbf{y}\}$ where, $\mathbf{p}$ is the component of $\mathbf{y}$ in the neutral speech subspace as shown in Figure 6.1. The set of projections of stressed speech vector, $\mathbf{y}$, onto the set of codevectors $\{\mathbf{a}_n\}$, $1 \leq n \leq N$, can be seen as projection of $\mathbf{y}$ onto neutral speech subspace. These components contain speech specific information of stressed speech vector. The error between $\mathbf{y}$ and $\{\mathbf{P}_n\mathbf{y}\}$ is orthogonal to the neutral speech subspace as shown in Figure 6.1. In other word, the error between these two vectors represents the projection of $\mathbf{y}$ onto the orthogonal

subspace. The orthogonal vector can be represented as $\mathbf{e} = \{(\mathbf{y} - \mathbf{P}_n\mathbf{y})\}$. Due to orthogonal assumption between speech and stress information, the orthogonal vectors will contain stress specific information of the stressed speech vector.

The projection of $\mathbf{y}$ onto the set of $\{\mathbf{a}_n\}$, $1 \leq n \leq N$ produces N projected vectors in neutral speech subspace and corresponding N orthogonal vectors in the orthogonal subspace. The stressed speech subspace can be decomposed into the speech ($\mathbf{S}$) and the stress ($\mathbf{X}$) subspaces. In this work, the stress vector $\mathbf{x}$ is decided by the orthogonal vector which has minimum length, as given in Eq. 6.6 and Eq. 6.7. The projected vector corresponding to that orthogonal vector is considered as speech vector as given in Eq.6.8.

$$\hat{n} = \arg\min_{n}[(\mathbf{y} - \mathbf{P}_n\mathbf{y})^T(\mathbf{y} - \mathbf{P}_n\mathbf{y})] \tag{6.6}$$

$$\mathbf{x} = (\mathbf{y} - \mathbf{P}_{\hat{n}}\mathbf{y}) \tag{6.7}$$

$$\mathbf{s} = \mathbf{P}_{\hat{n}}\mathbf{y} \tag{6.8}$$

6.2 Analysis of Speech and Stress Subspaces

The decomposition of stressed speech vector onto the speech and stress vectors is performed under the assumption that the stress information is orthogonal to the speech information. According to this assumption, the speech information should be present in the speech subspace, and the stress subspace should contain the stress information. The speech information in the stress subspace should be negligible and similarly stress information should be negligible in the speech subspace. This section verifies this assumption experimentally by using speech and stress recognition techniques. The underlying principle of this study is to investigate how effectively the stress and the speech information are separated.

In this work, $MFCC$ feature of the speech signal under stressed condition is represented as stressed speech vector. An angry speech vector of 20 msec segment of vowel /a/ is considered to demonstrate speech and stress vectors of the angry speech vector. The codevectors of vowel /a/ are developed by quantizing approximately 5000 frames of vowel /a/ from the neutral

speech. These codevectors, $\{\mathbf{a}_n\}$ are linear combination of neutral speech vectors of vowel /a/. Therefore, these vectors lies in the subspace of $\mathbf{R}^N$ and this subspace can be called as neutral speech subspace. In this work, 32 codevectors are developed using VQ technique [70]. The projection of angry speech vector onto the neutral speech subspace, $\{\mathbf{a}_n\}$, $1 \leq n \leq 32$ produces 32 projected vectors and corresponding 32 orthogonal vectors, as described in Section 6.1. The stress vector is considered as that orthogonal vector, which has minimum length among 32 orthogonal vectors. The projected vector corresponding to that stress vector is considered as a speech vector. The angry speech vector and its estimated speech and stress vectors are shown in Figure 6.2. In this figure, neutral speech vector is shown as that vector among the 32 neutral speech vectors where stress vector has minimum length. Figure shows that the variation pattern of estimated speech vector of the angry speech vector is approximately the

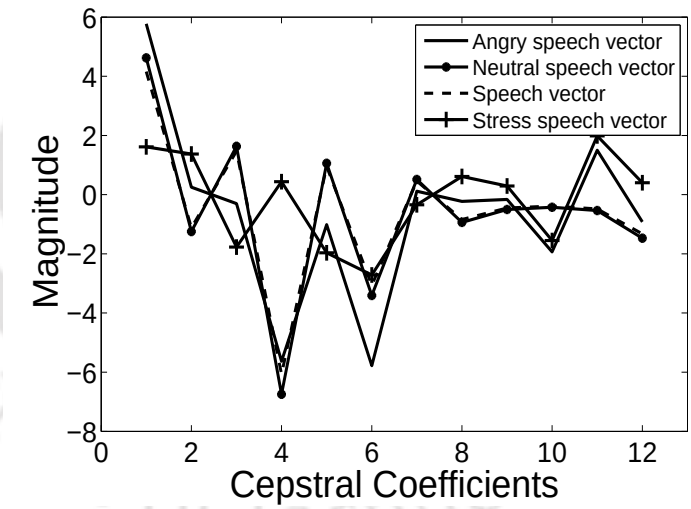

Figure 6.2: Analysis of speech and stress vectors of the angry speech vector

same as the neutral speech vector. This observation infers that the speech vector can capture the properties of neutral speech vector. On the other hand, the stress vector is the difference between the stressed speech vector and the speech vector. Therefore, it contains stress specific information. This section investigates the speech and stress information in speech and stress subspaces under the following two assumptions.

(i) When word and speaker information are present during projection which is speaker de-

pendent projection.

(ii) When word and speaker information are not present during projection which is speaker independent projection.

6.2.1 Speaker Dependent Projection

This subsection investigates speech and stress information in the speech and stress vectors under the assumption that the word and the speaker information are present during projection. To evaluate the speech and stress information, speech and stress recognizers are considered. The speech signal is segmented into a number of frames with the length of 160 samples and frame rate of 80 samples. $MFCC$ features of each of the frames of an utterance is considered as vectors of that utterance. A dataset of 30 word vocabulary is considered under neutral, angry, sad and Lombard conditions from four speakers (two males and two females) for this analysis. In this chapter, training is done using 1^{st} session of database and testing is always performed on 2^{nd} session of database.

6.2.1.1 Analysis of Speech Information

In order to provide speaker information during projection, speaker dependent neutral speech subspace is developed. The neutral speech subspace of a word is developed by considering all the utterances of that word taken from a single speaker. The neutral speech subspace of a word contains 32 vectors. The projections of stressed speech vectors, $\{\mathbf{y}\}$ of a given word onto the neutral speech subspace corresponding to that word form a set of speech vectors, $\{\mathbf{s}\}$ using Eq.6.8 and set of stress vectors, $\{\mathbf{x}\}$ using Eq.6.7. The same procedure is repeated for all the utterances present in the dataset for that speaker.

To evaluate the speech information in the speech and stress vectors, a speaker dependent HMM based speech recognizer is developed. The speaker dependent word model is developed using ten states and left to right transition. Each state has two mixture components. During recognition, the performances of stressed speech vectors $\{\mathbf{y}\}$ and corresponding speech vectors, $\{\mathbf{s}\}$ and stress vectors, $\{\mathbf{x}\}$ are computed. The recognition performances of speech and stress

vectors under different stressed conditions are investigated for two male and two female speakers. The recognition performances of these vectors for four speakers are shown in Figure 6.3. Figure

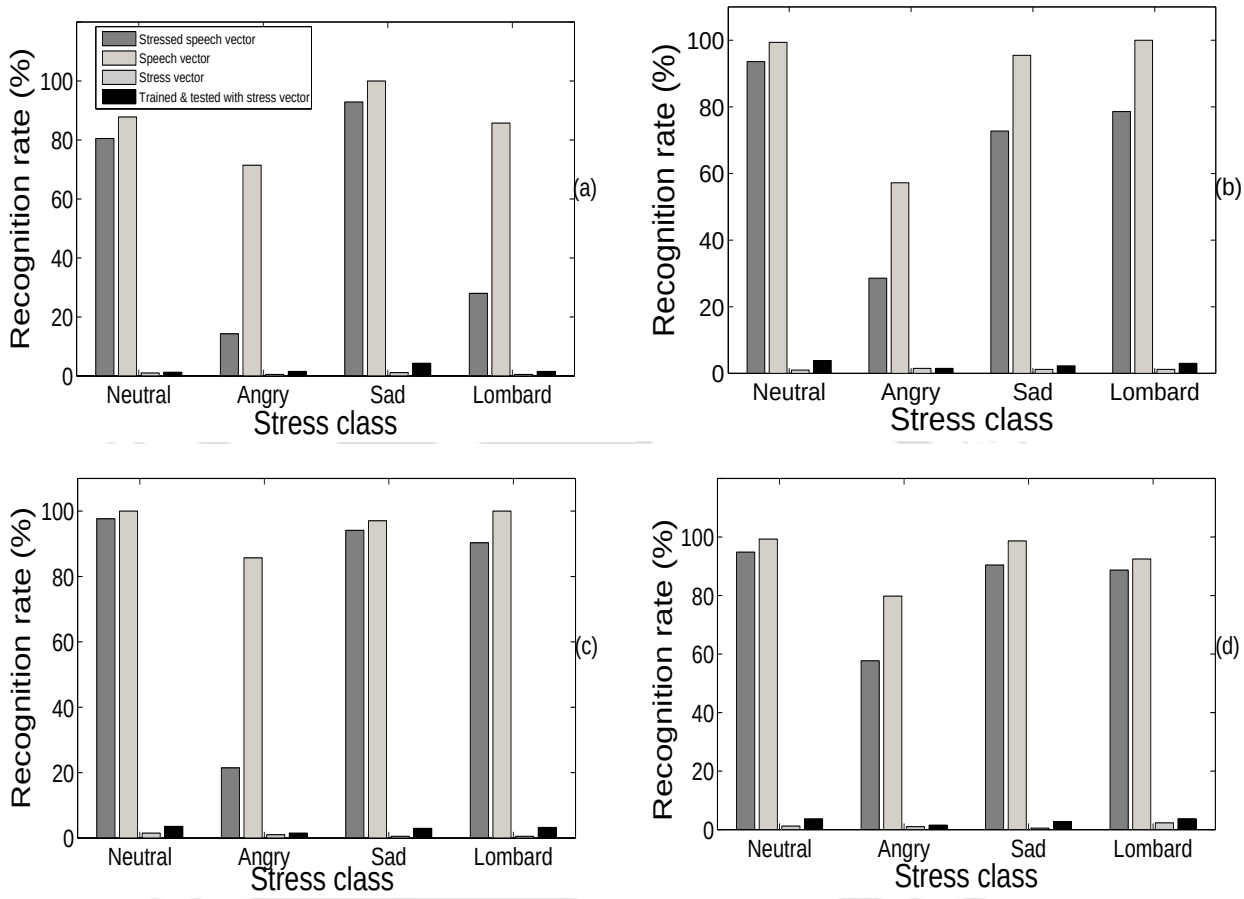

Figure 6.3: Performance of speech recognition of four speakers, (a) and (b) two female speakers, and (c)and (d) two male speakers

shows that the recognition performances of speech vectors, $\{\mathbf{s}\}$ under stressed conditions are higher than their stressed speech vectors, $\{\mathbf{y}\}$. The speech vector lies in the neutral speech subspace, and hence, it contains speech specific information. Alternatively, the recognition performances of stress vectors, $\{\mathbf{x}\}$ under stressed conditions are reduced significantly from that of stressed speech vectors, $\{\mathbf{y}\}$. From Figure 6.2, it can be observed that the dynamic range of stress vector is comparatively low from the neutral speech vectors. Therefore, the variation in dynamic range of stress vector may not give a fair comparison of the performances of $\{\mathbf{x}\}$ and $\{\mathbf{s}\}$. Hence, the speaker dependent speech recognizer is developed using stress vectors $\{\mathbf{x}\}$ and

the recognizer is also tested with stress vectors, $\{\mathbf{x}\}$. The recognition performances of stress vectors, $\{\mathbf{x}\}$ of four speakers are also shown in Figure 6.3. Figures show that under stressed condition, the recognition performance of stress vectors, $\{\mathbf{x}\}$ is still less than those of stressed speech vectors, $\{\mathbf{y}\}$ and speech vectors, $\{\mathbf{s}\}$. These observations are observed to be same for all speakers. This result infers that the speech information in the stress subspace is negligible.

6.2.1.2 Analysis of Stress Information

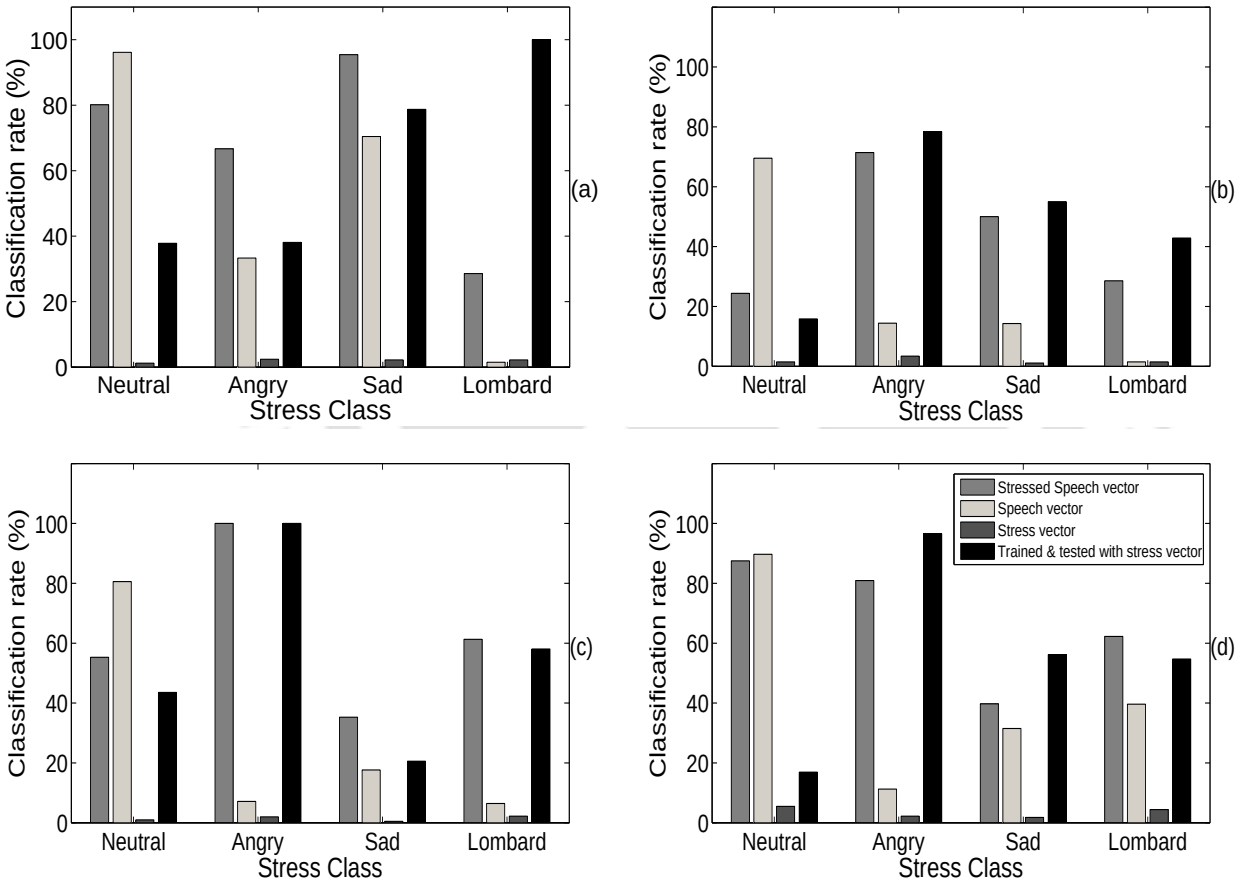

Figure 6.4: Classification rate of stress classifier of four speakers, (a) and (b) two female speakers, and (c)and (d) two male speakers

This subsection quantifies the level of stress in the speech and the stress subspaces using speaker dependent stress classifier. Stress classifiers are developed using HMM technique. The stressed speech vectors, $\{\mathbf{y}\}$ of all 30 words taken from a speaker are considered to model stress.

The stress model is developed using sixteen states and left to right transition with two mixture components per state. During testing, stressed speech vectors, $\{\mathbf{y}\}$ and their speech vectors, $\{\mathbf{s}\}$ and stress vectors, $\{\mathbf{x}\}$ of that speaker are tested under different stressed conditions. The classification rates of these vectors for four speakers are shown in Figure 6.4. From this figure, it can be observed that the classification rate of speech vector, $\{\mathbf{s}\}$ for angry, sad and Lombard conditions are reduced from that of $\{\mathbf{y}\}$. This observation infers that the speech vectors do not contain stress specific information. On the other hand, the classification rates of $\{\mathbf{x}\}$ degrade for angry, sad and Lombard conditions. The degradation in the classification rate of $\{\mathbf{x}\}$ may be due to the variation in dynamic range of stress vector as discussed in Subsection 6.2.1.1. Thus, the stress classifier is trained with stress vectors, $\{\mathbf{x}\}$ for a speaker and tested with stress vectors, $\{\mathbf{x}\}$ of that speaker under different stressed conditions. The classification rates of $\{\mathbf{x}\}$ of four speakers under different stressed conditions are shown in Figure 6.4. From the figures, it can be observed that the classification rates of stress vectors for angry, sad and Lombard stress classes are higher than that of speech vectors, $\{\mathbf{s}\}$. Also, the classification rate of stress vector, $\{\mathbf{x}\}$ for these stress classes are approximately same as that of stressed speech vectors, $\{\mathbf{y}\}$. This observation infers that $\{\mathbf{x}\}$ contains approximately the same stress information as stressed speech vector contained.

From Figure 6.3 and 6.4, it can be observed that the speech vectors contain speech information, and the stress information is approximately negligible in speech vectors. On the other hand, stress vectors contain stress information and the speech information is approximately negligible in stress vectors. This analysis infers that speech and stress subspaces obey orthogonality assumption.

6.2.2 Speaker Independent Projection

Previous subsection concluded that the speech and stress information are present in their respective subspaces, when the word and the speaker information are present during projection. In the practical scenario, the word and the speaker information are neither present during the projection nor to the recognizer. Therefore, this subsection investigates speech and stress

information in the speech and the stress vectors, when prior knowledge of word and speaker are not present during the projection as well as during the recognition. To evaluate speech information, speaker independent speech recognizer is considered. In this study, a dataset of 119 words vocabulary under neutral, angry, sad and Lombard conditions from fifteen speakers is considered.

6.2.2.1 Analysis of Speech Information

The projection of stressed speech vectors, $\{\mathbf{y}\}$ of an utterance onto the neutral speech subspace corresponding to a word form a set of speech vectors, $\{\mathbf{s}\}$ using Eq.6.8 and a set of stress vectors, $\{\mathbf{x}\}$ using Eq.6.7 of that stressed speech vectors. These sets of speech and stress vectors span the speech subspace, $(\mathbf{S})$ and stress subspace, $(\mathbf{X})$, respectively. In speech recognition task, 119 words, present in the database, can be assumed to form neutral speech subspaces $\{\mathbf{A}_w\}$, $1 \leq w \leq 119$ which are subspaces of $\mathbf{R}^N$. Thus, the projections of stressed speech vectors, $\{\mathbf{y}\}$ onto the neutral speech subspaces, $\{\mathbf{A}_w\}$ corresponding to 119 words form 119 speech subspaces, $\{\mathbf{S}_w\}$ and corresponding stress subspaces, $\{\mathbf{X}_w\}$. The stress subspace which has minimum average length among 119 stress subspaces is considered as stress subspace corresponding to that stressed speech vectors, $\{\mathbf{y}\}$. The speech subspace corresponding to this stress subspace is considered as speech subspace. The estimated speech and stress vectors which span to their respective subspaces are considered as speech and stress vectors of that stressed speech vectors. This analysis can also be seen as the speech subspace can be decided by that stress subspace which has minimum stress length. This study shows that the minimum stress length criterion has ability to decide the speech information in speech vectors [105]. In order to verify, how reliably this criterion measures the speech information in speech vector, $\{\mathbf{s}\}$, a VQ based speech recognizer is developed. In this study, the word model is considered with the neutral speech subspace of that word. The neutral speech subspace of a word is developed by considering approximately 40 utterance of that words taken from the fifteen speakers. Similarly, 119 neutral speech subspaces are developed using neutral speech taken from 119 words. The neutral speech subspace of a word contains 32 vectors. The speech information of given stressed speech vectors

are decided by that speech subspace which has minimum stress length. The performances of speech recognition using minimum stress length criterion under different stressed conditions are listed in Table 6.1. The performance of this criterion is compared with the Euclidean distance based speech recognition [106]. The Euclidean distance measures the length from $\{\mathbf{y}\}$ to $\mathbf{A}$. From the table, it can be observed that the recognition performance using minimum stress length criterion are higher than that of Euclidean distance for stressed conditions. This observation infers that the minimum stress length criterion has ability to decide the speech vector for stressed speech vectors.

Table 6.1: Performance of speech recognition using different distance measures under stressed condition

	Euclidean distance	Minimum stress length
Neutral	44.36	56.27
Angry	20.51	28.08
Sad	34.06	37.23
Lombard	29.73	34.22
Avg.Perform	32.16	38.95

In order to investigate the speech information in speech vectors, $\{\mathbf{s}\}$ decided by the minimum stress length criterion, speech recognition performance is evaluated using these speech vectors. In this study, the speech specific information in speech and stress vectors are investigated under different stressed conditions using speaker independent speech recognizer. During training, neutral speech is used to develop model for different words. In this work, the performances of these vectors are evaluated using VQ and HMM classifiers. For VQ classifier, 32 size of codebook is developed for a word. In HMM, the word model is developed using ten states, left to right transition and two mixture components per state. In this study, stressed speech vectors, $\{\mathbf{y}\}$ and corresponding estimated speech vectors, $\{\mathbf{s}\}$ and stress vectors, $\{\mathbf{x}\}$ are computed only for those utterances which are correctly recognized by the minimum stress length criterion given in Table 6.1. The recognition performances of these vectors using HMM and VQ classifiers are shown in Figure 6.5. Figure shows that the recognition performance of speech vectors,

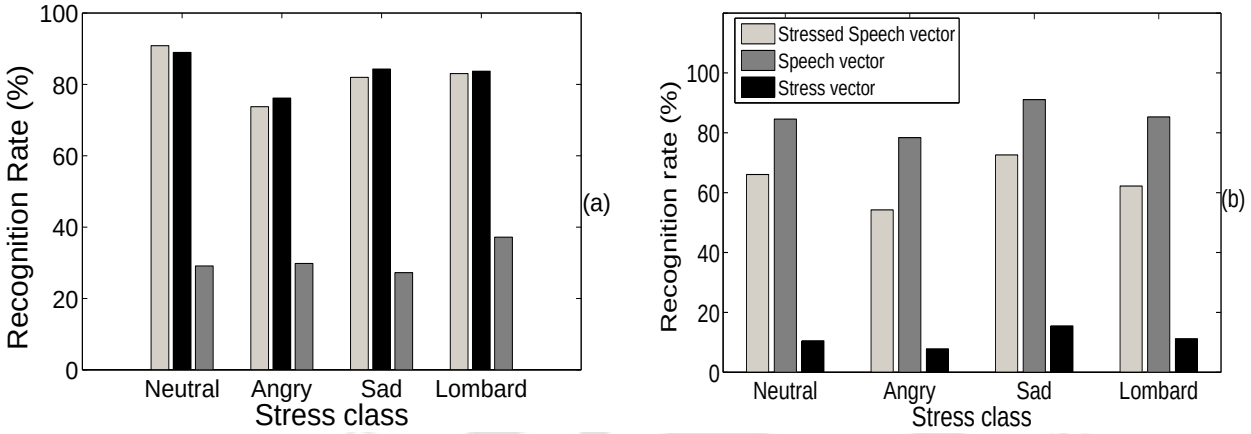

Figure 6.5: Performance of speech recognition using different models, (a) HMM model, and (b) VQ model

$\{\mathbf{s}\}$ under stressed conditions are higher than their stressed speech vectors $\{\mathbf{y}\}$. Alternatively, the recognition performances of stress vectors, $\{\mathbf{x}\}$ under stressed conditions are significantly reduced from that of $\{\mathbf{y}\}$. These observations are similar for both the classifiers.

6.2.2.2 Analysis of Stress Information

This section quantifies the level of stress present in speech and stress vectors by using stress

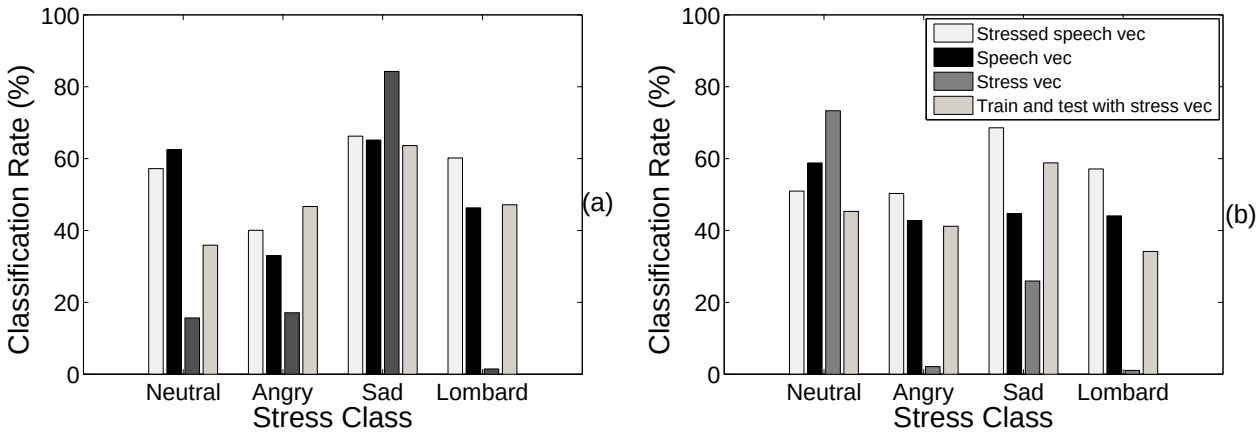

Figure 6.6: Performance of stress classification using different models, (a) HMM model and (b) VQ model

classifier. The stress classifiers are developed using VQ and HMM techniques. The stressed speech vectors, $\{\mathbf{y}\}$ of all the words from all the speakers are considered to model the stress.

For VQ classifier, 1024 size codebook is developed for a stress class. In HMM, the stress model is developed using sixteen states, left to right transition and two mixture components per state. During the testing, stressed speech vectors, $\{\mathbf{y}\}$ and their speech vectors, $\{\mathbf{s}\}$ and stress vectors, $\{\mathbf{x}\}$ are tested for those utterances which are correctly identified by the minimum stress length criterion. The classification rates of these vectors for different stress classes using HMM and VQ classifiers are shown in Figure 6.6. From the figures, it is observed that the classification rates of speech vector, $\{\mathbf{s}\}$ for angry, sad and Lombard classes are less than that of $\{\mathbf{y}\}$. From Figure 6.6(a), it is observed that the classification rates of stress vectors for angry, Lombard classes are higher than that of speech vectors, $\{\mathbf{s}\}$. The classification rate of stress vectors for sad class is approximately the same as that of speech vectors, $\{\mathbf{s}\}$. On the other hand, from Figure 6.6(b), it is observed that the classification rates of stress vectors for angry and Lombard classes are less than that of speech vectors, $\{\mathbf{s}\}$. The classification rate of stress vectors for sad class is higher than that of speech vectors, $\{\mathbf{s}\}$. These observations infer that the stress information is present in the stress vectors which is identified by one of the stress classifiers.

6.3 Subspace Projection Approach Based Stressed Speech Recognition

Section 6.2 concluded that the speech and stress specific information are present in their respective subspaces. Orthogonal assumption are first verified under the assumption that the word and speaker information are present during projection, as described in the Subsection 6.2.1.1. Secondly, the speech and stress information are verified in their respective subspaces only for those utterances, which are correctly recognized by the minimum stress length criteria, as described in the Subsection 6.2.2. However, in real scenario, neither the word nor the speaker information is presented to the recognizer. Hence, in this section, the estimated speech and stress vectors, as described in the Subsection 6.2.2, are used for speech recognition purpose to verify their respective information. The block diagram of speech recognition system using subspace projection based approach is shown in Figure 6.7. The speech and stress vectors are first separated from the stressed speech vectors using subspace projection based approach,

as given in Eq. 6.8 and Eq. 6.7. The estimated speech and stress vectors are then used for speech recognition. The estimated speech vectors based speech recognition requires two stages

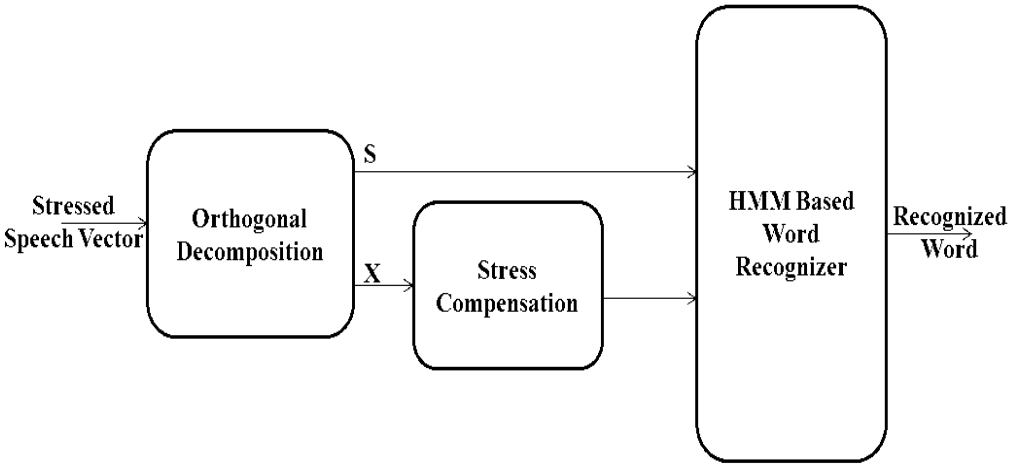

Figure 6.7: Block diagram of subspace projection based speech recognition

of recognition. The first stage estimates the speech vectors from the stressed speech vectors using minimum stress length criterion, as described in Section 6.2.1.1. In the second stage, the estimated speech vectors are used for speech recognition. Next, the stress information in estimated stress vector is verified by developing a stress compensation technique. In this technique, the estimated stress vectors are used to remove the effect of stress from the stressed speech vectors. The estimated stress vector contains stress information of that stressed speech vector. Therefore, the modified speech vectors of an utterance is computed by subtracting the average of estimated stress vectors of that utterance $\{\mathbf{x}\}$ from the stressed speech vectors $\{\mathbf{y}\}$ given as

$$\hat{\mathbf{y}} = \mathbf{y} - \mu_{\mathbf{x}} \tag{6.9}$$

where, $\hat{\mathbf{y}}$ is the modified speech vectors. The modified speech vectors are then used for speech recognition.

6.3.1 Error Characteristics of Speech Verification Using Subspace Projection Approach

The effectiveness of proposed speech and stress vectors based speech recognitions is evaluated

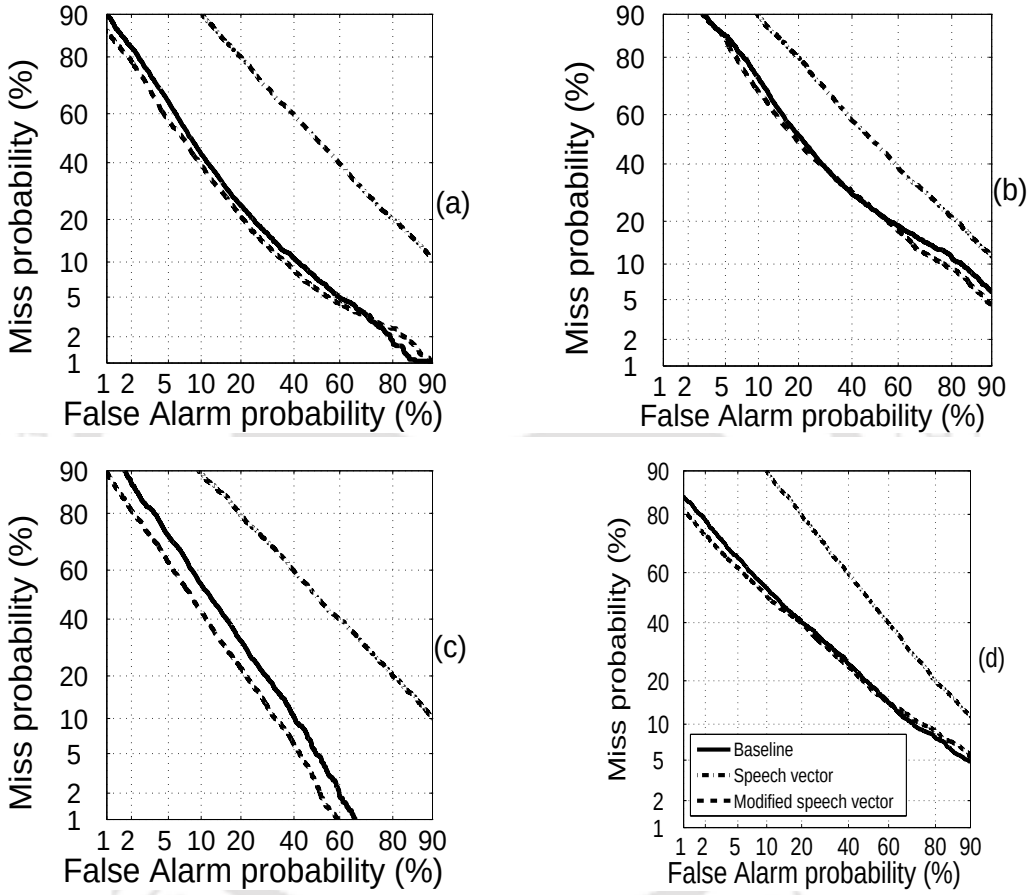

Figure 6.8: *DET* curves of speech and modified speech vectors based verification system for different stressed speech, (a) neutral, (b) angry , (c) sad, and (d) Lombard

using detection error trade-off (*DET*) curve. In this study, speech recognition is viewed as a verification task. A subset of 30 words vocabulary from fifteen speakers is considered for this analysis. In this study, the stressed speech vectors based verification system is considered as baseline. The word model is developed using neutral speech with ten states and left to right model. Each state has two mixture components. For testing, 1500 utterances are considered from each stress class. The genuine score can be considered as a likelihood value of a given word tested with given word model. Imposter scores can be considered as likelihood values

of a given word tested with other word models. Therefore, in this study, genuine scores are 1500×1 and imposter scores are 1500×29 for a given stressed condition. Similarly, the genuine and the imposter scores are computed for estimated speech and stress vectors. The estimated speech vectors based speech recognition is termed as speech vector based verification system and the speech recognition using modified speech vector, as given in Eq. 6.9, is termed as modified speech vector based verification system. The *DET* curves of speech subspace based verification under different stressed conditions are shown in Figure 6.8. From these figures, it is observed that the error rates of speech subspace based technique are significantly higher than error rates of baseline for all stressed conditions. Section 6.2.1.1 described that the minimum stress length criterion decides the speech vectors, $\{\mathbf{s}\}$ of stressed speech vectors, $\{\mathbf{y}\}$. Thus, any misidentification of this criterion forms incorrect speech subspace that may increase the error rate of this approach. This study infers that the speech subspace based verification requires explicit knowledge of word during projection. Therefore, this approach may not be an appropriate strategy for speech recognition. The *DET* curve of stress subspace based verification system is also evaluated for different stressed conditions and they are also shown in Figure 6.8. From these figures, it is observed that the error rates of stress subspace are lower than that of baseline for all stressed conditions. This observation infers that the stress vectors based compensation technique mainly eliminates stress biasing from the stressed speech vectors, and thus, this approach can be useful for speech recognition application.

6.3.2 Performance of Stressed Speech Recognition Using Subspace Projection Approach

In this section, the performance of speech recognition using subspace projection approach is evaluated. In subspace projection approach, the performance of the speech vector and modified speech vector based speech recognition are evaluated. The performances of speech and modified speech vectors based recognitions are tabulated in Table 6.2 with the name as speech vector and modified speech vector, respectively. The performances of these approaches are compared with the recognition performance of the *MFCC* feature. Here, the recognition performance of

the $MFCC$ feature is considered as baseline performance. Table shows that, the recognition performance of the $MFCC$ feature under neutral condition is 81.10%, whereas the average recognition performance of this feature under different stressed conditions is reduced to 59.82%. This result infers that the performance of the system degrades significantly under the stressed condition. The performance of speech vector based speech recognition reduces significantly

Table 6.2: Performance of subspace projection based speech recognition under stressed conditions

	Baseline	Speech vector	Modified speech vector
Neutral	81.10	61.77	80.96
Angry	42.58	30.37	44.05
Sad	59.89	43.09	63.91
Lombard	55.69	39.10	56.49
Average Performance	59.82	43.58	61.35

from that of baseline. The reduction in performance may be due to misidentification of {**s**} from {**y**} as described in Section 6.3.1. The recognition performance of modified speech vector for neutral speech slightly degrades, whereas the performances of other stressed conditions improve from the baseline feature. Further, the recognition performance improves significantly for sad speech. The recognition performances of this technique for angry, sad and Lombard speech are observed to be 44.05%, 63.91% and 56.49%, respectively, which are 1.47%, 4.02% and .8%, respectively, higher than those of baseline recognition performance. These results infer that although, the estimated speech vector contains speech information, as described in the Section 6.2, it depends on the identification rate of minimum stress length criterion. Therefore, the estimated speech vectors may not give correct speech information in the practical scenario. On the other hand, the estimated stress vector contains stress information, which is used to remove from the stressed speech vectors during compensation and therefore, the performance of speech recognizer using modified speech vectors is improved.

6.4 Speech Recognition Under Stressed Condition

Previous chapters demonstrated analysis of stress at different levels, namely, spectrum, sub-

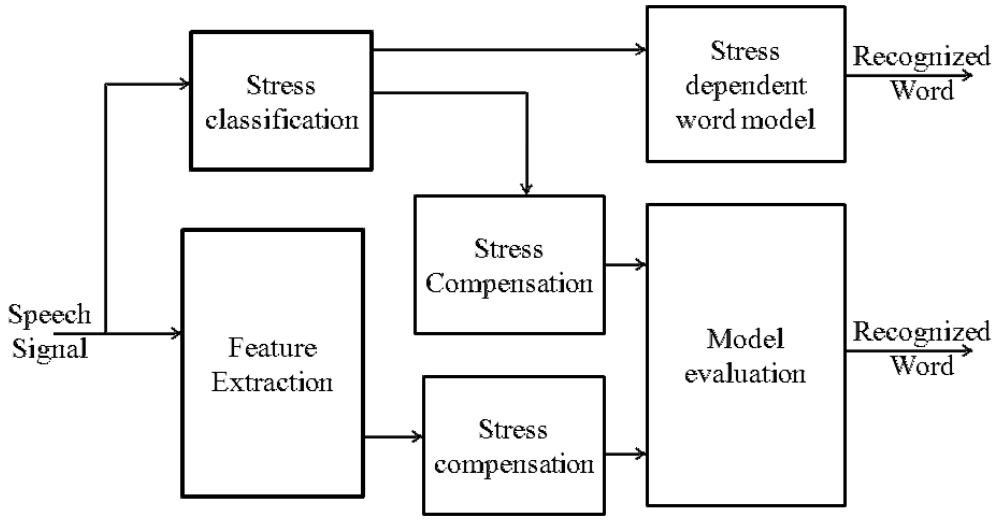

Figure 6.9: Block diagram of stressed speech recognition.

band and cepstral levels. Also, different approaches are demonstrated to model these stress information and used for speech recognition. This section presents the effectiveness of these stress analyses by comparing their speech recognition performances. The block diagram of speech recognition under stressed condition, based on our proposed analyses, is shown in Figure 6.9. The blocks of this recognizer is already discussed in Chapter 1. Chapter 4 shows effect of stress on the spectrum of the speech. Relative formant peak displacement based feature is proposed for characterization and classification of stressed speech. The effectiveness of proposed stress classification system is evaluated for speech recognition by using stress dependent speech recognizer and it is shown in Figure 6.9. Chapter 5 shows the effect of stress on subbands. The difference energy of subband is evaluated by developing stress compensation technique and the effectiveness of this technique is then evaluated using speech recognition and it is shown in Figure 6.9. The present chapter investigates the effect of the stress in cepstral coefficients using subspace projection based approach. The stress vector based compensation technique is also

shown in Figure 6.9. This section compares the recognition performances of these techniques

Table 6.3: Performances of proposed techniques for stressed speech recognition

	Baseline	Multistyle training	$RFD(log)$ -$MFCC$	Adaptive CMN	Difference energy	Subspace approach
Neutral	81.10	80.42	72.62	73.54	76.73	80.96
Angry	42.58	64.82	69.72	39.86	52.11	44.05
Sad	59.89	63.64	63.99	53.43	58.87	63.91
Lombard	55.69	75.14	75.19	52.14	65.63	56.49
Avg. Perform	59.82	71.14	70.38	54.74	63.34	61.35

under different stressed conditions and their performances are compared with the two existing techniques namely, multi-style training and adaptive cepstral mean normalization techniques. The recognition performances of these techniques are tabulated in Table 6.3. Different columns indicate different features or techniques of speech recognition under stressed condition. The multistyle training based approach is evaluated in Chapter 3 and the performance of this approach is tabulated in the 3^{rd} column of the Table 6.3. The stress dependent $MFCC$ feature based speech recognition evaluated in Chapter 4 and the performance of this technique is tabulated in the 4^{th} column of Table 6.3 as $RFD(Log)-MFCC$. By comparing the performances of multistyle training and $RFD(log)-MFCC$, it is observed that the recognition performance of neutral speech using $RFD(Log)-MFCC$ technique degrades from that of multistyle training. On the other hand, the recognition performances of other stressed conditions improve from those of multistyle training.

The adaptive mean cepstral compensation (CMN) technique assumed stress component additive to the speech component [11]. The recognition performance of this approach is also evaluated in Chapter 3 using SUSSC database and tabulated in the 5^{th} column of Table 6.3. The recognition performance of proposed $DSBE$ based compensation technique based speech recognizer tabulated in the 6^{th} column of Table 6.3. The recognition performance of proposed stress vector based compensation technique is tabulated in the 7^{th} column of Table 6.3. It can be observed from the table that the recognition performance of proposed stress vector based

compensation technique gives better performance than that of adaptive CMN technique. This observation infers that the estimation of stress, using orthogonal relation between speech and stress is more reliable than additive relation. By comparing the recognition performances of $DSBE$ and stress vectors based compensation techniques, it can be observed that, the average recognition performance of difference energy based approach is higher than that of stress vector based compensation technique. Although, $DSBE$ based approach gives improved performance compared to stress subspace and baseline, it requires additional stress information. On the other hand, stress subspace based compensation technique does not require neither stress information nor the information of level of stress present in the utterance.

6.5 Summary

This chapter evaluated a subspace projection approach for speech recognition. In this study, the stress information is assumed orthogonal to the speech information. Subspace projection approach is proposed to separate the stress and speech information from the stressed speech. The speech and stress specific information are verified using speech and stress recognitions. The speech and stress subspaces are further used for speech recognition purpose and the performance of these approaches are compared with existing techniques. Results show that the proposed stress subspace based compensation technique gives better performance compared to baseline and adaptive CMN techniques. On the other hand, speech subspace based approach does not perform effectively for speech recognition. Although, the performance of multistyle training is better compared to all other proposed techniques of speech recognition, it requires additional stressed speech database during training which is practically difficult to handle. On the other hand, the proposed technique does not require additional stress information during recognition. Therefore, this technique is computationally more easy to deploy for speech recognition under stressed conditions.

7

Conclusion

Contents

This thesis investigated the stress information in the spectral features of stressed speech. The investigation of stress information is further used for recognition of speech under stressed condition. A stressed speech database is developed to analyze the effect of stress in the speech feature. The stress and speech information are evaluated perceptually as well as by using automatic method. RFD is proposed for analysis of stress in the spectrum. To investigate the stress information in RFD, a stress classifier is developed, where RFD is considered as a feature for stress classification. Difference energy based feature is proposed for analysis of stress in the subbands. To investigate the insensitiveness of this feature to stress, difference energy based a stress compensation technique is proposed. In order to separate the stress and speech information in cepstral domain, a novel subspace projection based stress compensation technique is proposed.

In Chapter 3, the development and evaluation of a stressed speech database are presented. $SUSSC$ database is collected in Hindi, an Indian language, from fifteen non-professional speakers. A database of one hundred nineteen words is recorded for neutral, angry, sad, and Lombard conditions in two separate sessions. The stress and speech information present in $SUSSC$ database are validated by evaluating the level of stress and speech information present in the utterances. The perceptual validity of stress of this database is evaluated from the fifteen listeners. The ability of stress classification of the listeners is evaluated in terms of confusions between stresses. It was observed that most of the listeners confused neutral speech with sad speech and angry speech with Lombard speech. Sad speech is identified most of the time and Lombard speech is confused either with neutral or angry speech. It is also observed that some, among the set of listeners chosen, are poor at identifying stress. Also, some of the speakers, among the non-professional speakers are observed to be producing speech without proper stressed condition. After pruning out such listeners and speakers, the level of stress present in $SUSSC$ database is observed to be approximately similar to other existing database such as $SUSAS$. Also, the stress information is approximately the same in both the databases and hence, $SUSSC$ database can be used for speech recognition under stressed condition.

Content of the utterance of $SUSSC$ database is evaluated by listeners as well as automatic

speech recognition method. It is observed that the listener's perception of the content of speech is not effected by the stress present in the utterance. For automatic speech recognition, the $MFCC$ and the $LPCC$ features, along with their derivatives are evaluated under stressed condition. The recognition performance of the $MFCC$ and $LPCC$ features inferred that the performance of the speech recognition system degrades under stressed condition.

In Chapter 4, the stress information is investigated in the spectral tilt. The gross spectral tilt is first analyzed under different stressed conditions. It is observed that the gross spectral tilt has ability to discriminate the stress. This chapter also proposed local spectral tilt, which captures the relative variation in the formant peaks. The local slopes are termed as relative formant peak displacements (RFD). It is observed that the RFD values have ability to characterize the stress. The RFD values are concatenated with the gross spectral tilt, and they together are used as a feature for classification of stressed speech. The classification performance of the RFD feature is compared with the classification performance of the $MFCC$ feature. It is observed that the performance of the RFD feature is slightly less than the traditional $MFCC$ feature which shows that the RFD and the $MFCC$ features have approximately same stress discriminating capability. Further, the performance of smoothed log spectra derived RFD feature is higher than LP derived RFD feature. The RFD feature is combined with the $MFCC$ feature at the feature level, score level and rank level. The stress classification performances of all these combination techniques showed improvement from their individual feature performances. The proposed stress classification technique is then used for speech recognition purpose and the performance of stress dependent speech recognition shows improvement in the performance from the $MFCC$ feature based speech recognition.

In Chapter 5, the stress information is investigated in the subband energy. Statistical characteristics such as mean, variance and divergence of the subbands are investigated under different stressed conditions to analyze the stress information in the individual subband. Analyses show that the mean energy values of subbands for angry and Lombard speech are shifted to the higher level and mean energy values of subband for sad speech is shifted to the lower level than that for neutral speech. Also, the energy of higher subbands are deviated more than

the energy of lower subbands. In this study, difference energy of the subbands is proposed to measure the rate of migration of spectral energy. It is observed that the difference energy is less sensitive to stress than the other statistical characteristics such as mean, variance and divergence of subbands. Four stress compensation techniques are proposed to investigate the effectiveness of statistical and difference energy based analyses. These techniques are weighted mean, weighted variance, normalized divergence and difference energy. The effectiveness of these compensation techniques is evaluated using speech recognition system assuming that the recognition system has prior knowledge of stress class. The performance of stress compensation based speech recognition is compared with the $MFCC$ feature based speech recognition. The performances of weighted mean, normalized divergence and difference energy based compensation techniques showed improvement over $MFCC$ feature. However, the stress information is not known to the recognizer in practical scenario. The performances of weighted mean, normalized divergence and difference energy based compensation techniques are evaluated using rank level combination of $MFCC$ with RFD features. The performances of weighted mean, normalized divergence and difference energy based compensation techniques that the performance of the recognizer depends on the classification rate of the stress classifier. Thus, any failure of the stress classifier reduces the performance of the speech recognition system.

In Chapter 6, a subspace projection based approach is proposed to separate the speech and stress information from the stressed speech signal. In this technique, an orthogonal relation is assumed between the speech and the stress components. According to the orthogonal assumption, the speech information should be present in the speech subspace and the stress subspace should contain stress information. The speech information in the stress subspace should be negligible and similarly the stress information should be negligible in the speech subspace. Orthogonality assumption is verified experimentally by using speech and stress recognition techniques. The performances of speech recognition by using stressed speech vectors and by using their estimated speech and stress vectors are evaluated for different stressed conditions. This study showed that under stressed condition, the recognition performances of estimated speech vectors are higher than those of stressed speech vectors. On the other hand, under stressed conditions,

the speech recognition performances of estimated stress vectors are less than those of stressed speech vectors and estimated speech vectors. This study inferred that speech information is present in the speech subspace and the presence of speech information in stress subspace is negligible. The classification rate of stress classifier is evaluated using stressed speech vectors and their corresponding estimated speech and stress vectors for different stress classes. This study showed that for different stress classes, the stress classification rate of estimated speech vectors is less than those of estimated stress vectors. On the other hand, for different stress classes, the classification rate of estimated stress vectors is approximately the same as that of stressed speech vectors. These observations inferred that the stress vector contains approximately the same stress information as contained by stressed speech vector. The speech and stress recognition experiments concluded that the speech and stress specific information are present in their respective subspaces. The estimated speech and stress vectors are then used for speech recognition. The performance of estimated speech vectors based speech recognition is significantly less than that of the $MFCC$ feature. On the other hand, the performance of estimated stress vector based compensation technique is higher than the performance of the $MFCC$ feature. Also, the proposed stress compensation technique does not require additional stress information during recognition. Therefore, this technique is computationally more easy to deploy for speech recognition under stressed conditions.

Major contributions of the work reported in the thesis includes:

(i) A stressed speech database is developed.

(ii) The stress and speech information of stressed speech database is validated.

(iii) Relative formant peak displacement is proposed as a measure for spectral tilt.

(iv) Relative formant peak displacement based feature is proposed for stress classification.

(v) Difference energy based feature is proposed for stress compensation.

(vi) Subspace projection based approach is proposed for stress compensation.

7.1 Scope of future work

Some suggestions for future work are as follows:

(i) The stress information contained in the relative formant peak displacement can be used to remove the effect of the stress from the spectrum of the speech signal.

(ii) Glottal parameters show variation due to stress. The stress discriminating capability of these parameters can also be explored by developing stress classification system.

(iii) Difference energy of subband shows stress robust characteristic. It can be explored as a feature for speech recognition system.

(iv) Subspace projection based study considered orthogonal assumption. This study evaluates dot product between the speech and the stress information. Other inner products can also be explored between the speech and the stress information.

References

[1] B. J. Stanton, L. H. Jamieson, and G. D. Allen, "Acoustic-phonetic analysis of loud and Lombard speech in simulated cockpit conditions," in *Proc. Int. conf. on Acoust., Speech, and Signal Process.*, 1988, pp. 331–334.

[2] L. Rabiner and B. Juang, *Fundamentals of speech recognition.* New Jersey: Prentice Hall Inc., 1993.

[3] B. D. Womack and J. H. L. Hansen, "Classification of speech under stress using target driven features," *Speech Commun.*, vol. 20, pp. 131–150, Jun. 1996.

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

[5] R. A. W. Bladon and B. Lindblom, "Modeling the judgement of vowel quality differences," *J. Acoust. Soc. Amer.*, vol. 69, pp. 1414–1422, 1981.

[6] M. Ito, J. Tsuchinda, and M. Yano, "On the effectiveness of whole spectral shape for vowel perception," *J. Acoust. Soc. Amer.*, vol. 110, pp. 1141–1149, 2001.

[7] R. P. Lippmann, E. A. Mack, and D. . B. Paul, "Multi-style training for robust isolated -word speech recognition," in *Proc. ICASSP*, Apr. 1987, pp. 705–708.

[8] D. A. Cirns and J. H. L. Hansen, "Nonlinear analysis and classification of speech under stressed conditions," *J. Acoust. Soc. Amer.*, vol. 96, pp. 3392–3400, July 1994.

[9] S. Ramamohan and S. Dandapat, "Sinusoidal model-based analysis and classification of stressed speech," *IEEE Trans. Audio, Speech and Lang. Process.*, vol. 14, pp. 737–746, May 2006.

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

[11] Y. Chen, "Cepstral domain talker stress compensation for robust speech recognition," *IEEE Trans. on Acoust., Speech and Signal Process.*, vol. 36, pp. 433–439, April 1988.

[12] J. H. L. Hansen, "Morphological constrained feature enhancement with adaptive cepstral compensation (MCE-ACC) for speech recognition in noise and Lombard effect," *IEEE Trans. Speech Audio Process.*, vol. 2, pp. 598–614, Oct. 1994.

[13] G. Seshadri and B. Yegnanarayana, "Perceived loudness of speech based on the characteristics of glottal excitation source," *J. Acoust. Soc. Amer.*, vol. 129, no. 4, pp. 2061–2071, Oct. 2009.

[14] M. Kiefte and K. R. Kluender, "The relative importance of spectral tilt in monophthongs and diphthongs," *J. Acoust. Soc. Amer.*, vol. 110, pp. 1395–1404, 2001.

[15] S. G. Koolagudi and R. S. Krothapalli, "Two stage emotion recognition based on speaking rate," *Int. J. Speech Tech.*, vol. 14, no. 1, pp. 35–48, 2011.

[16] R. S. Bolia and R. E. Slyh, "Perception of stress and speaking style for selected elements of the SUSAS database," *Speech Commun.*, vol. 40, pp. 493–501, 2003.

[17] V. C. Tartter, H. Gomes, and E. Litwin, "Some acoustic effects of listening to noise on speech production," *J. Acoust. Soc. Amer.*, vol. 94, pp. 2437–2440, 1993.

[18] W. V. Summers, B. David, Pisoni, and H. Robert, "Effect of noise on speech production: Acoustic and perceptual analyses," *J. Acoust. Soc. Amer.*, vol. 84, pp. 917–928, 1988.

[19] S. Deligne and S. Dharanipragada, "A robust high accuracy speech recognition system for mobile applications," *IEEE Trans. Speech Audio Process.*, vol. 10, pp. 551–561, Nov. 2002.

[20] P. K. Rajasekaran and G. R. Doddington, "Speech recognition in the F-l6 cockpit using principal spectral components," in *Proc. Int. conf. on Acoust., Speech, and Signal Process. Dallas, Texas, USA*, vol. 10, Apr 1985, pp. 882–885.

[21] B. Beek, P. N. Edward, and D. C. Hodge, "An assessment of the technology of automatic speech recognition for military applications," *IEEE Trans. Acoust., Speech, and Signal Process.*, vol. 25, pp. 310–322, Aug 1977.

[22] C. J. Weinstein, "Opportunities for advanced speech processing in military computer-based systems," *Proc. IEEE*, vol. 79, pp. 1626–1641, Nov 1991.

[23] Y. Tsao and Chin-Hui, "An ensemble speaker and speaking environment modeling approach to robust speech recognition," *IEEE Trans. on Acoust., Speech and Signal Process.*, vol. 17, no. 5, pp. 1025–1037, July 2009.

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

[25] K. H. Davis, R. Biddulph, and S. Bailashek, "Automatic recognition of spoken digits," *J. Acoust. Soc. Amer.*, vol. 24, pp. 637–642, 1952.

[26] J. W. Forgie and C. D. Forgie, "Results obtained from a vowel recognition computer program," *J. Acoust. Soc. Amer.*, vol. 31, pp. 1480–1489, 1959.

[27] J. Makhoul, "Linear prediction: A tutorial review," *Proc. IEEE*, vol. 63, pp. 561–580, April 1975.

[28] H. Hermansky, "Perceptual linear prediction (PLP) analysis of speech," *J. Acoust. Soc. Amer.*, vol. 87, pp. 1738–1752, 1990.

[29] L. Rabiner and J. G. Wilpon, "Speaker independent recognition of isolated words using clustering techniques," *IEEE Trans. Acoust., Speech, Signal Process..*, vol. 27, pp. 336–349, 1979.

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

[31] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Trans. Acoust., Speech, and Signal Process.*, vol. 26, pp. 43–49, 1978.

[32] R. P. Lippmann, "An introduction to computing with neural nets," *IEEE ASSP Mag.*, vol. 4, pp. 4–22, April 1987.

[33] B. A. Dautrich, L. Rabiner, and T.B.Martin, "On the effect of varying filter bank parameters on isolated word recgonition," *IEEE Trans. Acoust., Speech, and Signal Process.*, vol. 31, pp. 793–806, 1983.

[34] A. P. Varga and R. K. Moore, "Hidden markov model decomposition of speech and noise," in *Proc. Int. conf. on Acoust., Speech, and Signal Process.*, 1990.

[35] M. Gales, "Robust continuous speech recognition using parallel model combination," *IEEE Trans. Acoust., Speech, and Signal Process.*, vol. 4, no. 5, pp. 352–359, Sep 1996.

[36] A. Acero and R. M. Stern, "Environmental robustness in automatic speech recognition," in *Proc. Int. conf. on Acoust., Speech, and Signal Process. Pittsburgh*, vol. 2, 1990, pp. 849–852.

[37] P. Krishnamoorthy and S. R. M. Prasanna, "Reverberant speech enhancement by temporal and spectral processing," *IEEE Trans. Speech, Audio and Lang. Process.*, vol. 17, pp. 253–266, Feb. 2009.

[38] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech and Lang.*, vol. 9, pp. 171–185, 1995.

[39] S. Furui, "Speech-to-text and speech-to-speech summarization of spontaneous speech," *IEEE Trans. Speech Audio Process.*, vol. 12, pp. 401–408, 2004.

[40] ——, "Recent progress in corpus based spontaneous speech recognition," *IEICE Trans. Inf. Syst.*, vol. 3, pp. 366–375, 2005.

[41] H. Soltau, B. Kingsbury, L. Mangu, D. Povey, G. Saon, and G. Zweig, "The IBM 2004 conversational telephony system for rich transcription," in *Proc. Int. conf. on Acoust., Speech, and Signal Process.*, vol. 1, 2005, pp. 18–23.

[42] Y. Liu, E. Shriberg, A. Stolcke, B. Peskin, J. Ang, D. Hillard, M. Ostendorf, M. Tomalin, P. Woodland, and M. Harper, "Structural metadata research in the ears program," in *Proc. Int. conf. on Acoust., Speech, and Signal Process.*, 2005.

[43] Y. Ichikawa, N. A., and K. Nakata, "Evaluation of various 'parameter sets in spoken digits recognition," *IEEE Trans. Audio and ElectroAcoust.*, vol. 21, pp. 202–209, June 1973.

[44] S. Furui, "Speaker independent isolated word recognition using dynamic features of speech spectrum," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 34, pp. 52–59, 1986.

[45] H. Hermansky and N. Morgan, "Rasta processing of speech," *IEEE Trans. on Acoust., Speech and Signal Process.*, vol. 2, no. 4, pp. 587–589, Oct 1994.

[46] 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 Lang.*, vol. 2, pp. 143–157, 1987.

[47] T. K. Vintsyuk, "Speech discrimination by dynamic programming," *Kibernetika*, vol. 4, pp. 81–88, 1968.

[48] Y. Tohkura, "A weighted cepstral distance measure for speech recognition," *IEEE Trans. on Acoust., Speech and Signal Process.*, vol. 35, no. 10, pp. 1414–1422, Oct 1987.

[49] F. Itakura, "Minimum prediction residual applied to speech recognition," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 23, pp. 67–72, 1975.

[50] C.H.Lee, C.H.Lin, and B. Juang, "A study on speaker adaptation of the parameters of continuous density hidden markov models," *IEEE Trans. Signal Process.*, vol. 39, no. 4, pp. 806–841, Apr. 1991.

[51] R. P. Lippmann, E. A. Martin, and B.Paul, "Multistyle training for robust speech recognition under stress," *J. Acoust. Soc. Amer.*, vol. 79, pp. S95–S95, April 1986.

[52] S. L. C. Busso and S. Narayanan, "Analysis of emotionally salient aspects of fundamental frequency for emotion detection," *IEEE Trans. Audio, Speech and Lang. Process.*, vol. 17, pp. 582–596, May 2009.

[53] J. H. L. Hansen and S. E. Bou-Ghazale, "Robust speech recognition training via duration and spectral -based stress token generation," *IEEE Trans. Speech Audio Process.*, vol. 3, pp. 415–421, Sept 1995.

[54] ——, "HMM-based stressed speech modeling with application to improved synthesis and recognition of isolated speech under stress," *IEEE Trans. Speech Audio Process.*, vol. 4, pp. 201–216, May 1998.

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

[56] S. Casale, A. Russo, and S. Serrano, "Multistyle classification of speech under stress using feature subset selection based on genetic algorithms," *Speech Commun.*, vol. 49, pp. 801–810, April 2007.

[57] E. Väyrynen, J. Toivanen, and T. Seppänen, "Classification of emotion in spoken finnish using vowel-length segments: Increasing reliability with a fusion technique," *Speech Commun.*, vol. 53, pp. 269–282, March 2011.

[58] Y. Xiao, J. Takatoshi, M. Chiyomi, K. Norihide, and T. Kazuya, "Classification of stressed speech using physical parameters derived from two-mass model," in *Proc. INTERSPEECH*, 2012.

[59] M. Afify, Y. Gong, and J. P. Haton, "A general additive and convolutive bias compnesation approach applied to noisy Lombard speech recognition," *IEEE Trans. Acoust., Speech, and Signal Process.*, vol. 6, pp. 524–538, 1998.

[60] J. H. L. Hansen and 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, Sept. 1995.

[61] R. Sarikaya and J. N. Gowdy, "Subband based classification of speech under stress," in *ICSSAP-1998*, 1998.

[62] G. S. Raja, "Feature analysis and compensation for speaker recognition under stressed condition," Ph.D. dissertation, Indian Institute of Technology Guwahati, Department of ECE , Guwahati, India, 2007.

[63] E. D. Cowie, N. Campbell, R. Cowie, and P. Roach, "Emotional speech: Towards a new generation of databases," *Speech Commun.*, vol. 40, pp. 33–60, 2003.

[64] D. Ververidis and C. Kotropoulos, "Emotional speech recognition: Resources, features, and methods," *Speech Commun.*, vol. 48, pp. 1162–1181, 2006.

[65] T. Seppänen, J. Toivanen, and E. Väyrynen, "Media team speech corpus: A first large finnish emotional speech database," in *Proc. of the 15th International Congress of Phonetic Sciences*, 2003.

[66] H. Steeneken and J. Hansen, "Speech under stress conditions: Overview of the effect on speech production and on system performance," in *Proc. Int. conf. on Acoust., Speech, and Signal Process., Phoenix, Arizona,*, vol. 4, mar 1999, pp. 2079–2082.

[67] J.H.L. Hansen, C.Swall, A.J.South, R.k.Moore, H.Steeneken, E.J. Cupples, T. Anderson, C.R.A. Vloeberghs, I. Trancoso, and P.Verlinde, "The impact of speech under 'stress' on military speech technology," *published by NATO Research and Technology organization RTOTR-10, AC/323(IST)TP/5 IST/TG-01*, 2000.

[68] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss., "A database of german emotional speech," in *Proc. INTERSPEECH*, 2005.

[69] R. Vergin and D. O. Shaughnessy, "Pre-emphasis and speech recognition," *CCECU, IEEE*, pp. 1062–1065, 1995.

[70] Y. Linde, A. Buzo, and R. M. Gray, "An introduction for vector quantizer design," *IEEE Trans. on Commun.*, vol. 28, pp. 84–95, 1980.

[71] F. Jelinek, "Continuous speech recognition by statistical method," *Proc. IEEE*, vol. 64, pp. 532–536, Apr 1976.

[72] G. F. Banks and L. W. Hoaglin, "An experimental study of duration characteristics of voice during the expression of emotion," *Speech Monographs*, vol. 8, pp. 85–90, 1941.

[73] G. Fairbanks and L. W. Hoaglin, "An experimental study of pitch characteristics of voice during the expression of emotion," *Speech Monographs*, vol. 6, pp. 87–104, 1939.

[74] Williams and Stevens, "Emotions and speech: Some acoustical correlates," *J. Acoust. Soc. Amer.*, vol. 52, pp. 1238–1249, Mar 1972.

[75] J. H. Hansen and S. Patil, "Speech under stress: Analysis, modeling and recognition," in *Speaker Classification I*, C. Müller, Ed., Berlin, Heidelberg, 2007, pp. 108–137.

[76] B. Heuft, T. Portele, and M. Rauth, "Emotions in time domain synthesis," in *Proc. Int. Conf. Spoken Lang. Processing*, vol. 3, 1996, pp. 1974–1977.

[77] A. Iida, N. Campbell, F. Higuchi, and M. Yasumura, "A corpus-based speech synthesis system with emotion," *Speech Commun.*, vol. 40, pp. 161–187, 2003.

[78] I. R. Murray and J. L. Arnott, "Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion," *J. Acoust. Soc. Amer.*, vol. 93, pp. 1097–1108, 1993.

[79] K. E. Cummings and M. A. Clements, "Analysis of glottal waveforms across stress styles," in *Proc. Int. conf. on Acoust., Speech, and Signal Process.*, 1990, pp. 369–372.

[80] K. W. Godin, T. Hasan, and J. H. L. Hansen, "Glottal waveform analysis of physical task stress speech," in *Proc. INTERSPEECH*, 2012.

[81] J. H. Hansen, "Evaluation of acoustic correlates of speech under stress for robust speech recognition," in *Proc. Int. conf. on Acoust., Speech, and Signal Process.*, 1989, pp. 31–32.

[82] S. Yildirim, M. Bulut, C. Lee, A. Kazemzadeh, Z. Deng, S. Lee, and S. Narayanan, "An acoustic study of emotions expressed in speech," in *Proc. Int. Conf. on Spoken Lang. Process.*, vol. 1, 2004, pp. 2193–2196.

[83] D. Klatt and L. C. Klatt, "Analysis, synthesis, and perception of voice quality variations among female and male talkers," *J. Acoust. Soc. Amer.*, vol. 87, pp. 820–857, 1990.

[84] Y. Lu and M. Cooke, "The contribution of changes in F0 and spectral tilt to increased intelligibility of speech produced in noise," *Speech Commun.*, vol. 51, pp. 1253–1262, 2009.

[85] B. J. Stanton, L. H. Jamieson, and G. D. Allen, "Acousticphonetic analysis of normal, loud, and Lombard speech in simulated cockpit conditions," *J. Acoust. Soc. Amer.*, vol. 84, pp. 115–115, 1988.

[86] T. Nwe, S. Foo, and L. D. Silva, "Speech emotion recognition using hidden Markov models," *Speech Comm.*, vol. 41, pp. 603–623, 2003.

[87] I. S. Engbert and A. V. Hansen, "Design, recording and verification of a danish emotional speech database," *Eurospeech 97, Rhodes, Greece,*, 1997.

[88] S. Steidl, "Automatic classification of emotion-related user states in spontaneous childrens speech," Ph.D. dissertation, Erlangen, 2009.

[89] B. Schuller, S. Steidl, and A. Batliner, "The interspeech 2009 emotion challenge," in *Proc. INTERSPEECH*, 2009.

[90] D. Ververidis and C. Kotropoulos, "A review of emotional speech databases," in *Proc. Panhellenic Conference on Informatics*, 2003, pp. 1–23.

[91] Sainz, Saratxaga, and Navas, "Subjective evaluation of an emotional speech database for basque," in *Sixth International Lang. Resources and Evaluation (LREC'08)*, 2008.

[92] S. T. Jovičić, Z. Kašić, M. orević, and M. Rajković, "Serbian emotional speech database: design, processing and evaluation," in *Conf. Speech and Computer St.Petersburg,Russia*, no. 9th, 2004.

[93] A. Iida, N. Campbell, and M. Yasumura, "A corpus-based speech synthesis system with emotion," *Speech Commun.*, vol. 40, pp. 161–187, 2003.

[94] B. Yegnenarayana and K. S. R. Murthy, "Event based instantaneous fundamental frequency estimation from speech signals," *IEEE Trans. Audio, Speech and Lang. Process.*, vol. 16, no. 8, pp. 1602–1613, Nov 2008.

[95] B. A. Hanson and T. Applebaum, "Robust speaker-independent word recognition using instantaneous dynamic and acceleration features: Experimnets with Lombard and noisy speech," in *Proc. Int. conf. on Acoust., Speech, and Signal Process.*, 1990, pp. 857–860.

[96] J. H. Hansen, "Evaluation of acoustic correlates of speech under stress for robust speech recognition," in *Proc. Bioeng. Conf. IEEE*, 1989, pp. 31–32.

[97] S. Kullback, "Information theory and statistics," *Dover, New York.*, 1968.

[98] H. Yeganeh, S.M. Ahadi, S.M. Mirrezaie, and A. Ziaei, "Weighting of mel sub-bands based on SNR/ entropy for robust ASR," in *Proc. Signal Process. and Inform. Tech.*, 2008, pp. 292–296.

[99] Y. Ephraim and H. L. V. Trees, "A signal subspace approach for speech enhancement," *IEEE Trans. Acoust., Speech, and Signal Process.*, vol. 3, no. 4, pp. 251–266, July 1995.

[100] A. W. Tan, M. Rao, and B. D. Sagar, "A signal subspace approach for speech modelling and classification," *Speech Commun.*, vol. 87, pp. 500–508, 2007.

[101] H. Lev-Ari and Y. Ephraim, "Extension of the signal subspace speech enhancement approach to colored noise," *IEEE Signal Process. Lett.*, vol. 10, pp. 104–106, 2003.

[102] J. Huang and Y. Zhao, "Energy-constrained signal subspace method for speech enhancement and recognition," *IEEE Signal Process. Lett.*, vol. 4, pp. 283–285, 1997.

[103] K. Y. Su and C. H. Lee, "Speech recognition using weighted hmm and subspace projection approaches," *IEEE Trans. Speech Audio Process.*, vol. 2, pp. 69–79, 1994.

[104] G. Strang, *Linear algebra and its applications*, 4th ed. Thomson Brooks/Cole, 2008.

[105] D. Mansour and B. H. Juang, "A family of distortion measures based upon projection operation for robust speech recognition," *IEEE Trans. on Acoustic, Speech and Signal Proc.*, vol. 37, pp. 1659–1671, 1989.

[106] B. A. Carlson and M. A. Clements, "A projection-based likelihood measure for speech recognition in noise," *IEEE Trans. Speech and Audio Process.*, vol. 2, pp. 97–102, 1994.

List of Publications

Journal Publication

1. Sumitra Shukla, S. Dandapat and S R Mahadeva Prasanna, "Spectral slope based analysis and classification of stressed speech" *Int J Speech Tech.*, vol. 11, Issue 3, pp. 245-258, 2011.

Conference Publications

1. Sumitra Shukla, S R Mahadeva Prasanna and S. Dandapat, "Stressed Speech Processing: Human vs Automatic in Non-professional Speakers Scenario" *IEEE Proc. NCC 2011*, Bangalore, India.

2. Sumitra Shukla, S R Mahadeva Prasanna and S. Dandapat, "Speech recognition under stressed condition" *IEEE Proc. NCC 2009*, Guwahati, India.

3. Sumitra Shukla, S. Dandapat, and S R Mahadeva Prasanna," Subspace Projection Based Analysis of Speech Under Stressed Condition", *IEEE Proc. WICT 2012*, Trivandrum, India.