Speech Subspace Modelling With Speaker Adaptation for Stress Normalization

Bhanu Priya

Speech Subspace Modelling With Speaker Adaptation for Stress Normalization

A

Thesis submitted

for the award of the degree of

DOCTOR OF PHILOSOPHY

By

Bhanu Priya

DEPARTMENT OF ELECTRONICS AND ELECTRICAL ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY GUWAHATI

GUWAHATI - 781 039, ASSAM, INDIA

APRIL 2018

This thesis is dedicated to my

Parents

and

Husband

for their love, support and encouragement

Certificate

This is to certify that the thesis entitled "**Speech Subspace Modelling With Speaker Adaptation for Stress Normalization**", submitted by **Bhanu Priya** (11610228), 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 my supervision and guidance. The thesis has fulfilled all requirements as per the regulations of the Institute and in our 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.

Date:

Place: Guwahati.

Dr. Samarendra Dandapat
Professor
Department of Electronics and Electrical Engineering
Indian Institute of Technology Guwahati
Guwahati - 781 039, Assam
India.

Acknowledgements

I feel it is as a great privilege in expressing my whole hearted and deep sense of gratitude to my thesis supervisor Prof. Samarendra Dandapat for his guidance, help and encouragement throughout my research work. I am highly grateful to him for patiently checking all my manuscripts and thesis. This thesis would not have been possible without their bounteous effort. He has always been a kind and motivating advisor and introduced me to the field of speech signal processing.

I would like to express my special thanks to Prof. Rohit Sinha, chairman of my doctoral committee for showing keen interest to the subject matter. I am highly obliged to him for giving kind support and valuable suggestions. I greatly admire their attitude towards research, creative thinking, hard work, dedication and honesty in work. These are the great source of inspiration for me in all my endeavors. I owe my profound gratitude to him for his supports in all respects.

My sincere thanks are due to my doctoral committee members Prof. S. R. Mahadeva Prasanna and Prof. Harshal B. Nemade for their support, encouragement and suggestions rendered during my research work. I am highly obliged to all the faculty members who have taught me various courses and provided me with active help and support, whenever needed.

I would like to acknowledge the support shown by my colleagues Dr. Anurag Singh, Dr. Sibasankar Padhy, Mr. Jiss J. Nallikuzhy and Dr. Nagaraj Adiga, who were with me during tenure of my research. I would like to thanks Dr. Haris B.C, Dr. Deepak K.T and Dr. Syed Shahnawazuddin for valuable discussions that we had time to time. I would like to thank Mr. Ajay Kumar Maddirala for the help provided during my programming stage. I acknowledge the help and my thanks to Dr. Mayank Naresh Agarwal. My sincere gratitude to all the research scholars of Electro-medical and Speech Technology Laboratory who work with me and it is a great research support and help. I would also like to express my thanks to the academicians at the Electronics and Electrical Department of IIT Guwahati for their support and understanding during the initial years of my Ph. D. studies. I am also grateful to the technical and the non-technical staff members of the department for their assistance in carrying out various tasks associated with this research work. I would also like to thank my friend Shekha Rai, Khaing Thin Zar and Shilpa Budhkar for supporting me and making my stay in IITG memorable.

My deepest gratitude goes to my parents for their unending love and support. Their continuous encouragement have enabled me to sail through the most difficult times during my studies. My

greatest debt of gratitude is owed to them, whose hard work and sacrifices gave me the opportunities I have today. Special thanks to my mother-in-law and father-in-law for their strong support. I would like to thank my sisters and my brother who have always motivated me during the course of my study. Finally, it is the love, affection, moral support and sacrifice of my husband '"Shubham", which has made this research a success. I sincerely acknowledge the help and support provided by him and I am extremely grateful. This dissertation is dedicated to my parents and my husband.

This acknowledgment will be incomplete without the mention of the Almighty Lord who has always showered His divine blessings and unceasing love and grace upon me and has protected me against all odd circumstances.

Bhanu Priya

Abstract

This thesis work is an investigation on the normalization of stress information for the effective processing of stressed speech. Speakers change the speech production system to communicate the information about the adverse environmental factors and to retain the intelligibility of speech signals. Any diversifications in the environmental condition from the normal or neutral state lead to an adverse condition and it is referred as the stress condition. The speech signal produced under stress condition by any modification in the speech production system is called as the stressed speech. The speech produced under normal or neutral condition is generally referred to as the neutral speech. Stress induces a large acoustic mismatch between the different speech units of neutral and stressed speech. These mismatched properties severely affect various real life applications. Thus, there is an essential need of stress normalization, that can reduce the acoustic mismatch between the neutral and the stressed speech and help the users with a better robust practical application. The present thesis aims at developing robust and computationally efficient algorithms to normalize the stress information.

First, novel linear and non-linear subspace modelling approaches are proposed to reduce the acoustic mismatch between the neutral and the stressed speech signals. The linear characteristic of stressed speech has been studied on the linear subspace. The linear subspace is modelled by exploiting an orthogonal projection and linear transformation techniques. The non-linearity between the speech and the stress information has been investigated on the non-linear data space by exploring the subspace projection through the non-linear transformation using the polynomial function. The results show that, the non-linear subspace modelling using the polynomial function of specific order is very effective for normalizing the stress information compared to the linear subspace modelling techniques. Secondly, an effective stress normalization method has been developed by investigating the changes in the vocal-tract system under stress condition in the Gaussian-

subspace. The acoustic mismatch between the vocal-tract system parameters of neutral and stressed speech is reduced by the subspace projection onto a common Gaussian-subspace, which consists of vocal-tract system parameters of neutral speech utterances. In this study, the proposed subspace projection is accomplished using the posterior probability information, which extracts the posteriorgram features. The synthesis of neutral and stressed speech signals using their estimated posteriorgram features corresponding to their vocal-tract system parameters has been observed very effective in reducing the acoustic mismatch between them. In the third approach, we have further investigated the deviation in the vocal-tract system under stress condition by exploring a novel subspace modelling technique in the sparse domain. In dictionary learning framework, two types of dictionaries have been proposed: the invariable size global dictionary and the utterance-specific adaptive dictionary. The invariable size global dictionary is learned using the well known K-SVD algorithm. The utterance-specific adaptive dictionary incorporates the information about the duration parameter of speech utterance, which is modeled using the K-nearest-neighbour (K-NN) algorithm. Both the neutral and the stressed speech are synthesized using their corresponding estimated vocal-tract system parameters and they are considered as the speech signals with the characteristics similar to the neutral speech. The experimental observations illustrate that, the sparse representation over the utterance-specific adaptive dictionary effectively reduces the acoustic variations between the vocal-tract system parameters of neutral and stressed speech signals with the significant improvement in stressed speech recognition performances compared to the conventional case and the case of using invariable size global dictionary.

Most of the methods reported in the literature to study the stressed speech are mainly based on the investigation of the changes in the speech production system under stress condition. It is observed that, variations in the anatomical and the physiological characteristics associated with different speakers under stress condition create a large acoustic mismatch between the neutral and the stressed speech. Hence, there is a need to develop robust methods which can normalize the speaker variabilities in the presence of stress. In order to mitigate the effect of such variabilities, at first, the heteroscedastic linear discriminant analysis (HLDA)- and the linear discriminant analysis (LDA)-based low-rank

subspace projections have been explored in the maximum likelihood linear transformation (MLLT)-based semi-tied adaptation technique. This helps in reducing the dimension as well as the correlation parameter of the feature- and the model-space. After that, the feature-space maximum-likelihood (fMLLR) transformations are generated for the training and the test speech utterances in the speaker adaptive training (SAT) mode to reduce the speaker variability. The effectiveness of proposed stress normalization methods are evaluated on three distinct frameworks namely: the stressed speech recognition, the visual analysis and the error analysis, respectively. The speech recognition for the stressed speech has been accomplished over the speaker dependent (SD) automatic speech recognition (ASR) systems employing acoustic models based on Gaussian mixture model (GMM), subspace Gaussian mixture model (SGMM) and deep neural network (DNN). In the visual analysis, we have studied the variations in the air pressure that human auditory system are able to perceive as sound and the changes in the spectral distributions with respect to the time by interpreting the waveform and the spectrogram, respectively. The error analysis measures the relative entropy between the Gaussian-subspaces by exploiting the Kullback Leibler (KL) divergence metric.

Keywords: Stressed Speech, Stress Normalization, Orthogonal Projection, Singular Value Decomposition, Non-linear Transformation, Polynomial Function, Posteriorgram Representation, Sparse Representation, Speech Recognition, Visual Analysis, Error Analysis.

Contents

List of Figures

List of Tables

List of Acronyms

ANN	: Artificial neural network
ASR	: Automatic speech recognition
CCA	: Canonical correlation analysis
CDHMM	: Continuous density hidden Markov model
CMVN	: Cepstral variance normalization
DaLSR	: Domain-Adaptive least-squares regression
DEAP	: Database for emotion analysis using physiological signals
DNN	: Deep neural network
EEG	: Electroencephalogram
EM	: Expectation-maximization
EMO-DB	: Database of German emotional speech
EP	: Emotion profile
ExpoLog	: Exponential-logarithmic
EWSC-HMM	: Error weighted semi-coupled hidden Markov model
FA	: Factor analysis
FIR	: Finite impulse response
fMLLR	: Feature space maximum-likelihood linear regression
GMM	: Gaussian Mixture Model
GOSP	: Generalized OSP
HCI	: Human-computer interaction
HMI	: Human-machine interaction
HMM	: Hidden Markov model
HLDA	: Heteroscedastic linear discriminant analysis
HRI	: Human-robot interaction

HTK	: Hidden Markov model toolkit
ISLSR	: Incomplete sparse least square regression
KL	: Kullback leibler
K-NN	: K-nearest-neighbour
LDA	: Linear discriminant analysis
LFPC	: Log-frequency power coefficient
LP	: Linear prediction
LPC	: Linear prediction coefficient
LPCC	: Linear prediction cepstral coefficient
MFA	: Mixture of factor analysis
ML	: Maximum likelihood
MMI	: Maximum mutual information
MFCC	: Mel-frequency cepstral coefficient
MLLT	: Maximum likelihood linear transformation
MOA	: Manner of articulation
MSE	: Margin scaling using the exponential function
MSH	: Margin scaling using the Hamming loss function
MSN	: Margin scaling using the linear function
NI	: No improvement
OMP	: Orthogonal matching pursuit
OSALPC	: One-sided autocorrelation linear prediction
OSP	: Orthogonal subspace projection
PLP	: Perceptual linear prediction
POA	: Place of articulation
RASTA-PLP	: Relative spectral PLP
SAL	: Sensitive artificial listener
SAT	: Speaker adaptive training
SC-HMM	: Semi-coupled HMM
SD	: Speaker dependent
SFM	: Spectral flatness measure

SGMM	: Subspace Gaussian mixture model
SI	: Speaker independent
STFT	: Short term Fourier transform
STP	: Short term processing
SUSAS	: Speech under simulated and actual Stress
SUSE	: Speech under simulated emotion
SUSSC	: Speech under simulated stress condition
SVD	: Singular value decomposition
T-ROT	: Thinking robot
TEO	: Teager energy operator
TEO-CB-Auto-Env	: TEO autocorrelation envelope area
UBM	: Universal background model
VTLN	: Vocal-Tract length normalization
VQ	: Vector quantization
WER	: Word error rate
WFBA	: Weighted filter-bank analysis

List of Symbols

a_p	: Linear prediction coefficient
a_{f_p}	: Linear prediction coefficient for the speech sample belongs to frame index f
$\tilde{a}_{f_p}$	: Estimated linear prediction coefficient for the speech sample belongs to frame index f
$\mathbf{a}_f$	: Set of linear prediction coefficients for the speech of frame index f, called as the LP-vect
$\boldsymbol{\alpha}_f$	: Posteriorgram features corresponding to the LP vector $\mathbf{a}_f$
$\mathbf{A}$	: Matrix containing LP vectors of observed speech utterance
$\tilde{\mathbf{A}}$	: Matrix containing estimated LP vectors of observed speech utterance
$\mathbf{c}_k$	: k^{th} row of cofactor of current estimate of HLDA matrix $\mathbf{H}$
$\mathbf{C}^w$	: Codebook matrix for the feature vectors of neutral speech utterances belong to word w
$\mathbf{D}$	: Dictionary
D	: Number of atoms in dictionary
$D_{\text{KL}}(f(x), g(x))$	: Kullback Leibler (KL) divergence between the probability density functions $f(x)$ and $g(x)$
$e(n)$	: Residual error
$e_f(n)$	: Residual error of the speech sample $s_f(n)$
E_c^w	: Average value of $\{\mathbf{v}_{f\ c}^{w}\}_{f=1}^{F}$
f	: Frame index
F	: Number of frames of observed speech utterance
$\mathbf{f}_f^w$	: Feature vector of neutral speech utterance belong to word w
f_{Mel}	: Mel-frequency
f_{Linear}	: Physical frequency
G	: Number of Gaussian densities in Gaussian mixture model
$\mathbf{h}_k$	: k^{th} row of HLDA matrix $\mathbf{H}$
$\mathbf{h}_s$	: Specific vector parameters of SGMM-based acoustic model
$\mathbf{H}$	: HLDA transformation matrix

$\mathbf{H}_L$	: Matrix containing first L rows of HLDA matrix **H**
$\mathbf{H}_{K-L}$	: Matrix containing last $(K-L)$ rows of HLDA matrix **H**
$\mathbf{H}_s$	: Global shared matrix of SGMM-based acoustic model
I	: Number of Gaussian densities in GMM- and SGMM-based acoustic models
K	: Number of elements in feature vector
L	: Dimension of decorrelated subspace
L	: LDA transformation matrix
$\mathbf{L}_L$	: Matrix containing first L rows of LDA matrix **L**
$\mathbf{L}_{C-L}$	: Matrix containing last $(C-L)$ rows of LDA matrix **L**
$\mathbf{m}_g^{\alpha}$	: Mean vector of g^{th} Gaussian density of GMM developed using the LP vectors of training data
$\mathbf{M}^w$	: Mean vectors of GMM corresponding to word w of training data
M	: Mean vectors of GMMs developed using all words of training data
N	: Total frames present in the training data
N^w	: Number of feature vectors of neutral speech utterances belong to the word w
N_c	: Total frames belong to the class c of training data
P	: Order of LP analysis
P	: Transformation matrix
P_g	: Posterior probability corresponding to the gth-Gaussian density for a given $\mathbf{a}_f$
p_n	: Order of polynomial function
$p_s(\mathbf{x}_t)$	: Continuous probability density function for the emission of speech features in the state s
Q	: Number of triangular bandpass filters in filter-bank
R	: Size of codebook
S	: Number of states of GMM- and SGMM-based acoustic models
$s(n)$	: Speech sample
$s_f(n)$	: Speech sample belongs to frame index f
$\tilde{s}_f(n)$	: Synthesized speech sample corresponding to $s_f(n)$
T	: Transpose operation
T	: Covariance matrix of training data
T_0	: Sparsity constraint
$\mathbf{U}_j^w$	: Matrix containing the j bases of speech subspace corresponding to word w

$U_j^{w'}$	: Filtered speech subspace
$\mathbf{v}_f$	: Stressed speech feature vector
$\mathbf{v}_{f_c}^{w'}$	: Complement orthogonal projection component of stressed speech feature vector
$\mathbf{v}_{f_p}^{w'}$	: Orthogonal projection component of stressed speech feature vector
W	: Total words present in the training data
$\mathbf{W}_c$	: Within class covariance matrix of the training data of c^{th} class
$\mathbf{W}^{\varphi}$	: fMLLR transformation matrix corresponding to speaker φ
w_{si}	: Weight parameter of GMM- and SGMM-based acoustic models
ω_f	: Cell constituting K_n nearest-neighbor LP-vectors of neutral speech utterances for $\mathbf{a}_f$
$\mathbf{x}$	: Arbitrary feature vector of speech signal
$\mathbf{x}_t$	: Feature vector of speech signal corresponding to frame at time t
$\mathbf{x}_{t_C}$	: Time-spliced feature vector
$\mathbf{X}_A$	: Matrix containing sparse coded vectors corresponding to $\mathbf{A}$
$\mathbf{X}_Y$	: Matrix containing sparse coded vectors corresponding to $\mathbf{Y}$
$\mathbf{y}$	: Non-linearly transformed feature vector
$\mathbf{Y}$	: Matrix containing LP-vectors of neutral speech utterances
$\mathbf{Y}_f$	: Matrix containing K_n nearest-neighbor LP-vectors of neutral speech utterances for $\mathbf{a}_f$
Y_q	: Log-energy output of the q^{th} filter
$\mathbf{z}$	: Decorrelated feature vector
$\mathbf{z}_L$	: Vector containing first L elements of vector $\mathbf{z}$
$\mathbf{z}_{K-L}$	: Vector containing last $(K-L)$ elements of vector $\mathbf{z}$
$\mathbf{z}_{t_C}$	: Time-spliced decorrelated feature vector
$\mathbf{z}_{t_L}$	: Vector containing first L elements of vector $\mathbf{z}_{t_C}$
$\mathbf{z}_{t_{C-L}}$	: Vector containing last $(K-L)$ elements of vector $\mathbf{z}_{t_C}$
$\mathbf{z}_{t_L}^{\varphi}$	: Decorrelated features corresponding to speaker φ
$\hat{\mathbf{z}}_{t_L}^{\varphi}$	: fMLLR adapted decorrelated features from a particular speaker φ
φ	: Index for speaker
$\boldsymbol{\mu}_{si}$	: Mean vector of GMM- and SGMM-based acoustic models
$\boldsymbol{\Sigma}_{si}$	: Covariance matrix of GMM- and SGMM-based acoustic models
$\boldsymbol{\xi}_{t_L}^{\varphi}$	: Feature vector $\mathbf{z}_{t_L}^{\varphi}$ along with the unity element

1

Introduction

Contents

1. Introduction

This thesis documents our investigations on the acoustic mismatch between the speech signals produced under neutral and stress conditions. The speech signal is a wave of acoustic sound pressure [1–3]. The production of speech signal involves the movement of anatomical structures (lungs, trachea, larynx, pharyngeal cavity, oral cavity, nasal cavity etc.) of speech production system. The speaker modifies the speech production system to emphasize or de-emphasize the information about the adverse environmental factors and to retain the intelligibility of speech signal [4–8]. The emotional state, emergency condition, multitasking, high workload, fatigue, sleep deprivation, Donald Duck-effect, gravitational force, pathological conditions etc. are the causes for the adverse environmental condition, which induce stress. Speech produced under stress condition by any alteration of the speech production system is called as the stressed speech or speech under stress [1, 2, 4–7, 9, 10]. On the other hand, the speech produced under normal or neutral condition is generally referred to as the neutral speech. Stress influences the speech production system and it results in a large acoustic variation between the neutral and the stressed speech. This mismatch is mainly attributed to the diversity in the pitch, the formant frequencies, the intensity, the energy, the average phone duration, the speaking rate, the glottal parameters and the pronunciation [5, 6, 11]. Consequently, many real life, large scale laboratory and commercial applications particularly in the areas of text-to-speech synthesis, speech coding, education, military, security, detection of mental and physical health of pilot, information retrieval, voice dialing, banking transactions through voice network, database access service, language learning, reservation systems, voice mail, security control for confidential information access, remote access to computers, entertainment applications, which involve their interaction with machines exhibit highly degraded performances for the users under stress conditions [8, 9, 11–18]. The reduction in the acoustic mismatch between the neutral and the stressed speech, which is generally referred to as the stress normalization will effectively improve the performances of those tasks for the user under stress condition. The key motivation for researchers is the designing of signal processing and pattern matching algorithms without degrading in-depth quality and intelligibility of speech signals produced under both the neutral and the stress conditions, respectively.

The investigation on the changes in the characteristics of stressed speech can be exploited for the implementation of effective stress normalization algorithms. In literature, various techniques are developed in the feature domain to investigate the acoustic mismatch between the neutral and the stressed speech and are considered as the state-of-the-art for the research concerns involving the

stressed speech processing [19–26]. Furthermore, in recent years, great advances have been made in the modelling paradigm for the robust and the explicit representation of stressed speech units [27–33]. The primary objective of all the research studies is to develop the subspace corresponding to the stressed speech comprising the characteristics similar to the subspace created using the neutral speech. The subspace projection-based methods have received a great deal of attention over the past several years because of their good realization capability, existence of efficient algorithms and elementary computational properties for the investigation in the deviation in the properties of stressed speech [33–51]. Although stress normalization methods with promising results are proposed, there are still a number of fundamental problems, which have motivated this work. In this thesis, the normalization of stress-specific attributes has been dealt with three major contributions.

The first approach of the thesis investigates the novel subspace projection-based methods to reduce the acoustic mismatch between the neutral and the stressed speech. The proposed stress normalization techniques explore the subspace projection for investigating the linear and the non-linear characteristics between the speech and the stress information. The linear characteristic between the speech- and the stress-specific attributes has been exploited by an orthogonal projection-based subspace filtering approach. In order to address the non-linear characteristics, the discrepancy between the neutral and the stressed speech are reduced by the subspace projection onto the non-linear data space derived by exploring the non-linear transformation using the polynomial function.

The second approach is intended towards the normalization of stress information by investigating the modification in the vocal-tract system characteristics under stress condition. To develop an effective stress normalization algorithm, a novel subspace projection-based approach over the vocal-tract system parameters has been proposed. The vocal-tract system parameters for the neutral and the stressed speech are projected onto a common subspace. This common subspace has Gaussian features and it consists of vocal-tract system parameters of neutral speech data. The subspace projection is derived using the posterior probability information and it has resulted in the posteriorgram features. Both the neutral and the stressed speech are synthesized using their corresponding projected vocal-tract system parameters to reduce the acoustic mismatch between them. It is hypothesized that, the synthesis of neutral and stressed speech using their corresponding estimated vocal-tract system parameters helps in retaining the similar acoustic properties.

In the third approach, the acoustic mismatch between the vocal-tract parameters of neutral and

stressed speech are reduced by exploring the subspace projection in sparse domain. The proposed subspace projection is accomplished through the linear transformation using the over-complete linear models also called as the dictionary learned on the vocal-tract system parameters of neutral speech under which the observed signal can be sparsely coded using a few suitable atoms. In this work, two different learning mechanisms are proposed for an estimation of effective dictionary. The first learning mechanism exploits the well known K-SVD algorithm for the estimation of invariable size global dictionary. In addition to this, we have incorporated the information about the duration parameter of speech utterance, which is modeled by exploiting the K-nearest-neighbour (K-NN) algorithm-based non-parametric probability density estimation method to construct the utterance-specific adaptive dictionary. Both the neutral and the stressed speech are synthesized using their respective estimated vocal-tract system parameters to reduce the acoustic mismatch between them.

In addition to these studies, the presented work in this thesis is also to search for a subspace having dimension lower than that of the default dimension and consisting of speech information along with the suppressed speaker variability. The resulting low-rank features, in which speaker information is also normalized significantly reduce the acoustic mismatch as well as the computational complexity of the training and the test environments of the automatic speech recognition (ASR) system. In modelling paradigms, we have explored the speaker dependent (SD) ASR systems employing acoustic models based on Gaussian mixture model (GMM) [1], subspace Gaussian mixture model (SGMM) [52–54] and deep neural network (DNN) [55, 56].

The remainder of this Chapter is organized as follows: The overview of stress normalization is presented in Section 1.1. Section 1.2 contains the detailed analysis of stressed speech by studying the variation in the acoustic-space, the feature-space and the model-space under stress condition. In Section 1.3, the subspace projection-based approaches are discussed for analyzing the acoustic mismatch between the neutral and the stressed speech signals. Section 1.4, portrays the consequences of speaker variability in the stressed speech. The scope for the present investigations are discussed in Section 1.5. Finally, organization of the thesis is summarized in Section 1.6.

1.1 Overview of Stress Normalization

The speech signal is one of the most fundamental and natural mode of communication for the human beings [1–3]. In our everyday experience, stress conditions affect the speech production sys-

tem. Speech produced under stress condition constitutes the modified acoustic properties, when compared to the speech produced under neutral condition. The changes in the characteristics of speech severely degrades the performance of numerous aforementioned real life applications, which involve their interaction with machines possible for various contemporary tasks. All these practical applications require the development of robust and computationally efficient human-computer interaction (HCI) system that can involve the human beings under stress conditions [23, 24, 57–59]. The foremost requirement for the development of effective HCI system is to create the training and the test environments of speech recognition system having similar characteristics. The normalization of stress-specific attributes would significantly reduce the variance mismatch between the training and the test environments of speech recognition system trained on the neutral speech, when tested using the stressed speech. The stress normalization provides robust and precise representation of stressed speech patterns and it helps in accomplishing the user-affable real life applications.

Speaker produces the acoustic sound pressures through a series of neurological activities and the movements of anatomical structures [1–3]. The speech signal constitutes the series of these acoustic sound pressures and it carries the significant information about the phonetic, linguistic, speaker and environmental characteristics [60–62]. The frequency or the spectral contents of the speech signal produced from the time varying speech production system change continuously with time [1–3]. The speech signal exhibits the quasi-stationary characteristics. The quasi-stationary properties are not reliable for processing of speech signal using the existing signal processing tools, which are based on the assumption of time invariant system and time invariant signal, i.e. stationary signal [3, 63]. To make use of existing signal processing tools, the short term processing (STP) of speech signal is proposed, in which it is assumed that the speech signal conveys the stationary characteristics within a block of time duration of 10 to 30 msec. Speech signal is mainly categorized into the vowels, fricatives, affricates, nasals, diphthongs, liquid, glides and stops based on properties related to the place of articulation (POA) and the manner of articulation (MOA). Whereas, the nature of excitation source classifies the speech into the three classes namely: the voiced speech, the unvoiced speech and the silence region, respectively. The voiced speech is produced by the speech production system, when excited using the nearly periodic impulse sequences, generated from the periodic movement of the vocal fold. The periodicity associated with the voiced speech is referred to as the fundamental frequency of input excitation and is also called as the pitch frequency. Since, the voiced speech is pe-

riodic in nature, the spectrum of voiced speech consists of harmonics of the fundamental frequency. The random noise-like excitation source helps in exhaling the air out of lungs through the trachea without any interruption by the vocal folds and it results in the production of unvoiced speech. The silence region is obtained, when no excitation is given to the speech production system. The silence region between the succession of voiced and unvoiced speech is an important aspect for retaining the intelligibility of speech signal. Moreover, the anatomical and the physiological properties of the excitation source and the vocal-tract system varies for different speakers. Different speakers under the similar stress environmental condition produce the same speech utterances with varying intensity of stress levels. The high frequency values of pitch is observed in children's speech compared to the pitch frequency of adults. The investigation on the changes in the speech production system under stress condition is the primary objective of the researchers involved in the area of stress normalization. These investigation helps in studying the acoustic mismatch between the neural and the stressed speech for the developement of effective stress normalization techniques.

1.2 Assessment of Stressed Speech

In literature, several algorithms for processing of stressed speech with good performances have been reported. The normalization of stress-specific attributes helps in robust and computationally efficient processing of stressed speech. The development of effective stress normalization algorithm exhibits three major problems. The first problem associated with the investigation on the changes in the acoustic parameters of speech production system under stress condition. The second problem deals with the extraction of effective features, which constitute the phonetic information along with the suppressed stress-specific attributes. The third problem addresses the robust and the precise representation of stressed speech patterns in modelling paradigm. To address these problems, in this Chapter, the stressed speech is analyzed onto the three distinct but interconnected spaces namely: the acoustic-space, the feature-space and the model-space, respectively, as depicted in block diagram shown in Figure 1.1. The studies of the characteristics of stressed speech onto these spaces can help in effectively normalizing the stress-specific attributes. The following Subsections contain the details on the assessment of stressed speech onto the aforementioned spaces.

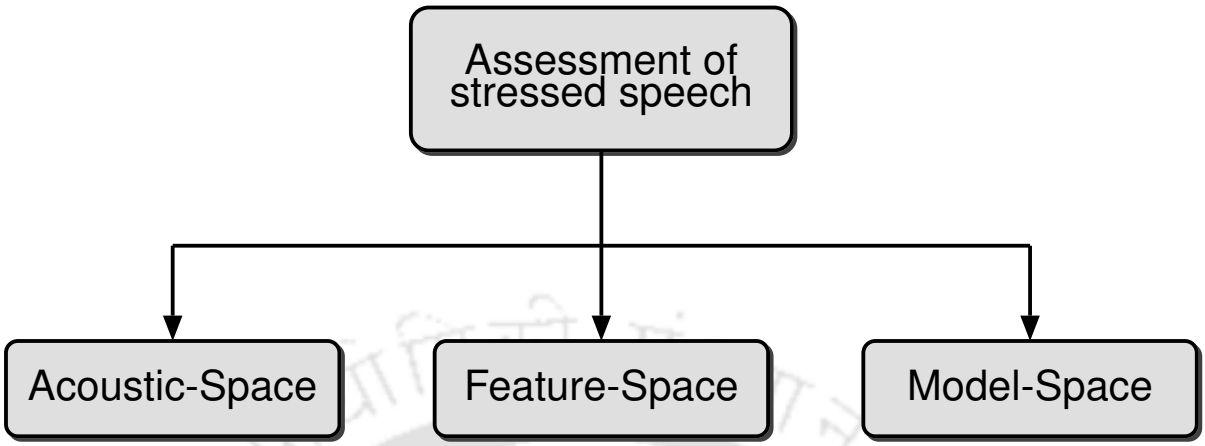


Figure 1.1: Assessment of stressed speech onto the acoustic-space, the feature-space and the model-space for the development of effective stress normalization techniques.

1.2.1 Assessment of Stressed Speech in the Acoustic-Space

In the presence of stress, the fundamental aspect of human communication manifests the significant information about the expression of stress and the ability to perceive such stress. The analysis of parameters of the speech production system that reflect the significant information about the stress can lead to informative direction for the investigation on the aforementioned acoustic mismatch between the neutral and the stressed speech signals. The following paragraphs describe the assessment of stressed speech onto the acoustic-space.

Investigation by Carl et al. [5, 64] is one of the foremost studies. They have done a detailed analysis on the variations in the speech production system parameters for the speech produced under anger, fear and sorrow conditions. The alteration in fundamental frequency, spectrum, temporal characteristics, precision of articulation and waveform regularity of successive glottal pulses were studied to reflect the emotional state of speakers. It was shown that, the fundamental frequency for the production of neutral speech varied smoothly and slowly with respect to the time. Whereas, the presence of stress leads to irregular pattern of fundamental frequency contour. The average value of fundamental frequency for the anger voice was noted to be high in comparison to the fundamental frequency for the speech produced under fear, sorrow and neutral conditions. The mean value of fundamental frequency for the speech uttered under sorrow condition was observed lower than the fundamental frequency for the neutral speech. In another work [60], the modifications in the duration

parameter were studied under the five emotion categories namely: contempt, fear, anger, grief and indifference, respectively. The modifications in the duration parameters were investigated for the word, the vowel, the semi-vowel, the consonant and the diphthong. Considerable degree of changes in the duration parameter were observed in the production of vowels and consonants under studied emotion classes. These alterations in duration parameters were considered as important aspects for the classification of emotion categories. Hansen and Patil [6] have investigated the variations in the pitch, the fundamental frequency, the duration, the intensity, the glottal pulse shaping and the vocal tract spectrum for the stressed speech. The increased value of duration was noted for the vowels produced under loud and angry circumstances in comparison to the neutral condition. In comparison to the duration of vowels spoken under neutral condition, decreased value of vowel duration was observed for the soft speech. The intensity of vowels has increased value for the speech produced under angry, soft and question conditions compared to the intensity of vowels for the neutral speech signal. In another study [65], alterations in the speech production system were investigated by studying the vocal expression and the perception. The fundamental frequency, the jitter and the shimmer were considered as the potential parameters for perceiving the information about emotion classes. In that work, it was found that, the fundamental frequency for the production of joy speech has increased value in comparison to that of the neutral speech. Philip et al. [62] have studied the fundamental frequency and the amplitude parameters for the transmission of emotional content under bored statement, confidential communication, expressing disbelief doubt, fear, happiness, objective question, objective statement and pompous statement categories.

The effect of noisy environment on speech production system was first studied by the French otorhino-laryngologist, Etienne Lombard [66]. The noisy environment also induces stress in the speech and it is generally referred as the lombard effect. The resulting speech produced under noisy environment is called as the lombard speech. The literature survey shows that, the investigation on parameters of speech production system is also effective for the analysis of lombard speech. The work reported in [61] investigates the changes in the acoustic parameters and the perceptual characteristics for the production of lombard speech uttered by the female and the male speakers. The significant degree of variation was noted for the pitch and the energy parameters in the production of speech by the female and the male speakers under the lombard effect. The results of that work indicate that, the changes in the characteristics of phoneme under lombard condition depends

on the context information. Hansen and Patil [6] have also studied the various acoustic parameters for the analysis of lombard speech. It was observed that, the intensity of vowel remained constant under lombard conditions. Whereas, the glottal pulses with steeper slopes and pointed corners were found under lombard condition in comparison to the neutral condition. The maximum variability in spectral amplitude was observed between 2 kHz to 4KHz frequency band for the lombard speech. In the work reported in [21], the spectral envelope was investigated for the classification of angry, sad and lombard stress categories. It was found that, frequency values for the high formants were affected more under stress condition than the formant frequencies under low frequency region. The spectral tilt comprising the gross energy information of spectral envelope was noted to have a flatten structure for the lombard speech compared to the spectral tilt for the neutral speech.

1.2.2 Assessment of Stressed Speech in the Feature-Space

This Section elaborates the detailed studies of various techniques used for analyzing the characteristics of stressed speech in the feature-space. The investigation of acoustic mismatch between the neutral and the stressed speech in the feature-space is considered as the state-of-the-art for the effective processing of stressed speech. In literature, numerous features related to the time, the frequency and the time-frequency domains have been suggested for the analysis of stressed speech. In the following paragraphs, we have presented the detailed studies on the various approaches explored in the feature-space for the assessment of stressed speech.

In literature, significant degree of variations are reported in the frequency domain for the stressed speech compared to the neutral speech [19–22,41]. Investigation by Bou-Ghazale and Hansen is one of the primary studies [19]. They have studied the variation in the characteristics of stressed speech in the individual frequency bands. In that work, it was shown that, the sensitivity of stress is inconsistent with respect to the frequency and they have introduced the Mel-frequency (M-MFCC) and the exponential-logarithmic (ExpoLog) scales. In addition to this, they have also proposed the one-sided autocorrelation linear prediction (OSALPC) and the cepstral information-based OSALPC features, which were found to be very effective for measuring the variation in the properties of stressed speech. In another work [22], the distribution of the spectral energy among the different sub-bands on the Mel-scale was analyzed using the Log-Frequency Power Coefficient (LFPC) features. The work reported in [20] has interpreted the set of harmonic sequences, called as the Fourier parameters for quantifying the perceptual contents of speech produced under emotional condition. To accomplish that,

the first- and the second-order differences of harmony features were derived for the speech emotion recognition and they were found to be effective for modelling the variation induced in speech due to emotions. Zhou et al. [44, 45] have studied that, the air propagates within the vocal-tract system in a non-linear trend under adverse stressful condition. The interpretation of Teager energy operator (TEO)-based processing as the mathematical model representation of non-linear airflow pattern reflects the variation in the airflow during the production of speech under stress condition. In that work, three types of TEO-based features are investigated namely: the TEO-FM-Var, the TEO-Auto-Env and the TEO-CB-Auto-Env. The TEO-FM-Var features are derived from the frequency modulated (FM) component of speech. The normalized TEO autocorrelation envelope (TEO-Auto-Env) measures the variation in excitation source within the uniform frequency bands of audio frequency range. Whereas, the TEO-CB-Auto-Env features, referred to as the critical band based TEO autocorrelation envelope are determined by fine partition of frequency bands onto the Mel-scale to reflect the variation in the fundamental frequency more pronouncedly under stress condition. The experimental observations presented shows that, TEO-based features lead to the development of efficient technique for the classification of speech signals produced under different stress conditions. In other work reported in [25], the airflow patterns in the physiological system model were studied by exploiting the two-mass model for the classification of stress classes. The airflow patterns were modeled using the physical parameters of two-mass model and reported to be very effective for classifying the stress classes. The sinusoidal features were shown to have attractive representation advantages in the spectral domain for stress classification [41]. The amplitude, the frequency and the phase components of sinusoidal representations were introduced for indexing the emotion classes. The experimental results illustrate that, the amplitude values in the low frequency region has better emotion classification capability than the amplitude values for the higher frequencies. In the work reported in [21], alterations in the properties of spectral envelope are studied. In that work, the slope of spectral envelope, called as the spectral tilt, which constitutes the gross energy information was used for measuring the variation in the spectral properties under stress condition. The spectral tilt values are quantified using the displacement of amplitude values corresponding to second, third and fourth formants with respect to the amplitude of the first formant. They were noted to be deviated under stress condition than the neutral condition. It was found that, the influence of stress is less on the first formant peak and it rises with the increasing value of formant frequencies. Deng et al. [67] have investigated the variation in the

phase spectrum for the classification of whispered speech. The characteristics of phase spectrum were extracted from the modified group delay and the all-pole group delay features. These features were found very effective for classifying the whispered speech in comparison to the non-whispered speech produced under different stress conditions. The experimental evaluation of the presented work indicates that, the physical parameters are very robust for recognizing the stress classes.

The prosodic characteristics and the source parameters of speech production system carry meaningful information about the stress classes [26, 68]. Sundberg et al. [26] have studied twelve source parameters of speech production system for the analysis of speech produced under emotional condition namely: equivalent sound level, alpha ratio, maximum flow declination rate, relative duration of the closed phase, level difference between the first and the second harmonics, difference between the average long term averaged spectrum level near mean and average level one octave higher, source flow pulse amplitude, normalized amplitude quotient, mean fundamental frequency, jitter, shimmer and harmonics to noise ratio. It was observed that, the physiological mechanism in each emotion class comprises the specific combination of these studied source parameters. In that work, the sad speech was represented by a low sub-glottal pressure, the weak glottal adduction and the lower fundamental frequency. The production of angry speech acquires the high sub-glottal pressure and the lower level of shimmer. Whereas, the joy speech was portrayed by high values of both the sub-glottal pressure and the fundamental frequency. In the work reported in [68], the prosodic, the spectral envelope and the voice quality features were studied for speech emotion recognition. It was found that, the prosodic properties constitute the precise information about the emotion and thus the features related to the variation in the prosodic characteristics have resulted in better classification accuracy in comparison to the spectral envelope-based features. Furthermore, the supra-segmental parametrization of these features was turned to the additional improvement in the accuracy of classification of speech produced under anger, boredom, disgust, fear, happiness, neutral, and sadness cases.

Several research works have been devoted for the study of lombard effect on the speech production system [69–72]. Varadarajan et al. [70] have done detail analysis on the impact of lombard onto the speech signals produced under different types and levels of noises using the duration, the energy histogram and the spectral tilt features. It was noted that, speaker pronounces the decreased value of duration for fricatives, diphthongs, nasals, semi-vowels, affricates, stops, silence and the increased value of duration for the vowels to emphasize the lombard effect. Furthermore, a decreased value

of spectral tilt has been observed to signify the high frequency components of the lombard speech. In another work [71], the lombard effect was normalized in the frequency and the cepstral domains in the unsupervised manner. In frequency domain, the spectral components of the short segment of speech were adapted in the neutral speech patterns. Whereas, the dynamics of cepstral coefficients were estimated using the quantile method in the cepstral domain. The proposed algorithm has resulted in better recognition performances with the severer deterioration in the WER in comparison to the WER obtained from the baseline recognition system. In the work reported in [72], the morphological constrained feature enhancement with adaptive cepstral compensation (MCE-ACC) algorithm was developed to suppress the lombard effect. To accomplish this algorithm, the source generator frameworks were established for the neutral and the lombard speech by employing the morphological feature constrained enhancement along with the stressed source compensation techniques. The HMM system developed using the proposed algorithm was found to be very effective and has resulted in the significant improvement in the recognition performances for the lombard speech.

1.2.3 Assessment of Stressed Speech in the Model-Space

There are growing research in the field of stressed speech by exploring an effective representation or modelling of different speech units of stressed speech. The robust and the precise optimization of model parameters for the stressed speech allows the speech researchers to investigate the deeper understanding into the characteristics of stressed speech. Recent studies have reported explicit modelling of anatomical and physiological changes during stress. The following paragraphs summarize the various modelling techniques developed in unsupervised, supervised and semi-supervised manners to optimize the model parameters for the assessment of stressed speech.

Most of the works reported in literature have optimized the Markov process developed in supervised manner for the robust modelling of stressed speech units [27–29]. Womack et al. [27] have modified the Markov process model to attain improved performances of tasks involving the stressed speech recognition and the stress classification. The Markov process model was modified by developing the multidimensional hidden Markov model, called as the $N-$channel HMM accomplished by allowing the transition between one-dimensional HMMs. The modification in the speech production system under stress condition were represented by the dimension of $N-$channel HMM. The experimental observations illustrate that, the $N-$channel HMM is a better representation of perceptually evoked stress in the sub-phoneme. It has resulted in improved performances for the recognition

and the classification of stressed speech. In another work reported in [28], the Markov process was studied for the human emotion recognition using the audio-visual bimodal signal. In the proposed approach, an error weighted semi-coupled hidden Markov model (EWSC-HMM) has been developed. The implementation of EWSC-HMM has utilized the integration of semi-coupled HMM (SC-HMM) onto the framework of state-based bimodal alignment and Bayesian classifier weighting scheme. The time dependent correlation between the audio and the video samples was modeled using the state-based bimodal alignment scheme of the SC-HMM. The classification strategy of EWSC-HMM has adopted the weighted Bayesian classification method. The infusion of correlation between the audio and the video streams into the EWSC-HMM was noted to contribute significant improvement in the recognition of emotion classes. Afify et al. [29] have proposed a novel technique in the model-space to compensate the lombard effect under different noisy conditions. The proposed algorithm has employed bias in the state level of continuous density hidden Markov model (CDHMM) in maximum likelihood framework. The bias models were developed to diminish the mismatch between the clean and the lombard speech. These bias models, when developed in the Mel-cepstral and the linear-spectral domains have been considered as the additive models. The parameters of bias model was further precisely optimized using the polynomial trend in the Mel-cepstral domain and it has provided an effective compensation of lombard effect.

In addition to the development of modelling technique in supervised learning framework, various learning algorithms have utilized the unsupervised and the semi-supervised learning mechanisms to estimate the model parameters for the effective modelling of speech produced under stress condition [23, 31, 32, 59, 73]. Zhang et al. [31] have proposed an affective computing of human emotion recognition from the speech data by exploring the novel data-cooperative learning method. The data-cooperative learning is based on the efficient sharing of labelled work between the low confidence value of human labelling and the high confidence value of the machine labelling. The proposed algorithm for the implementation of cooperative learning was developed in three different paradigms using the active learning and the semi-supervised learning methods. The first approach has utilized the active learning and the self-training for the execution of the single-view cooperative learning. In the second method, the mixed-view cooperative learning has been built, which combined the active learning and the co-training techniques. The third method addressed the co-active learning and the co-training for the implementation of the multi-view cooperative learning. The experimental results

of that work show that, in comparison to the single-view cooperative learning and the multi-view cooperative learning, the mixed-view cooperative learning algorithm has accomplished improved performance of speech emotion recognition. The work reported in [73], has addressed the emotion recognition in the missing data scenario. To overcome the inefficacy of the information from the missing data, the multi-modal ensemble-based system was developed. The ensemble process has been accomplished using the standard feature-level fusion and the decision-level fusion techniques. The proposed ensemble method was noted to be very effective for the recognition of emotion using voice, face, and gesture features in the missing data condition. Cowie et al. [23] have noted that, the paralinguistic features constitute the potent information about the stress and they were effectively extracted from the neural nets. It was also observed that, some of the speech correlates exhibit the similar characteristics under emotional condition. In another work [59], the emotion profiles (EPs), which contain the emotion-specific information were explored for the better performance of human-machine interaction (HMI) system, when operated by the users under stress condition. The estimation process of EPs has utilized the output of the four SVM classifiers trained in unsupervised framework. They were weighted by the estimate of the confidences of the respective assignments. The human expression was modelled using multiple probabilistic class label determined from the EPs and reported to be very promising for classifying the emotion labels. In a recent work [32], the domain-adaptive least-squares regression (DaLSR) model was proposed for the cross corpus speech emotion recognition. The DaLSR model employed the joint training of labelled training data and unlabelled target data. The proposed joint training has resulted in the reduction of variance mismatch induced due to the cross corpus and provided the improved emotion recognition performances.

1.3 Subspace Modeling Approaches for the Stressed Speech

The stress induces diversity in the acoustic parameters of speech production system. These diversities modify the properties of stressed speech. The investigations on the changes in the characteristics of stressed speech have included various signal processing algorithms in the acoustic-space, the feature-space and the model-space. The development of robust and computationally efficient algorithms is the foremost requirement for implementation of various practical applications in real life scenario. Among the numerous techniques explored in literature, the subspace projection-based technique is one of the most popular and widely accepted method in all fields of research. The

subspace projection method leads to an effective investigation on the subject in the framework of appealing realization capacity, efficient algorithms and simple computational aspects. Over the past several years, the subspace projection-based approaches have received a great deal of attention in the field of stressed speech processing.

In the area of stressed speech, the subspace projections are derived mainly by exploring the linear and the non-linear relationship between the speech and the stress information. Hansen et al. [69] have proposed the subspace projection onto the source generator framework. The source generator framework was created using spectral features in an iterative manner. The proposed iterative approach was found to be very effective for enhancement and equalization of the source generator. The source generator optimization method has resulted in the pronounced technique for capturing the variation in a production of speech under the emotional and noisy conditions. In another work [40], an orthogonal relationship between the speech- and the stress-specific attributes is assumed to separate these two information from the speech produced under stress condition. The stressed speech is orthogonally projected onto the speech subspace and was reported to be very effective for normalizing the stress information. In that work, the speech subspace was created by performing the K-means clustering [1, 2] over the neutral speech features. The experimental observations of that work manifest the importance of estimation of an effective speech subspace. In the work reported in [34], an orthogonal noise signal decomposition onto the framework of rank-one projection was proposed for the normalization of multi-microphone noise. The multi-microphone noise is effectively suppressed by maximizing the total variance of the coherent noise.

In last few years, the sparse representation and the compressive sensing are extensively used for recognition and classification of speech produced under stress condition. This provides a new direction for the stressed speech processing research [42, 47, 51, 74, 75]. The sparse representation using the over-complete matrices are reported to provide a viable alternative to the projection under the full-rank matrices [76–78]. In the work reported in [74], the dictionary is learned in a multiview supervised manner for multiview representation analysis of speech emotion recognition. In the work reported in [75], the incomplete sparse least square regression (ISLSR) model is used for speech emotion recognition. The ISLSR is based on the linear relationship between the speech feature and the corresponding emotion labels. Jia-Ching et al. [42] have proposed mapping of variance on non-uniform auditory scale-frequency in sparse domain for emotion verification. In another work [51],

mean parameters of HMM were exploited in sparse domain to improve the accuracy of speech recognition for the stressed speech. These studies manifest that, the iterative process of sparse representation results in the more precise and robust representation of stressed speech features, which, in turns, leads to an effective technique for analyzing the stressed speech.

Speech signal comprises mainly of phonetic and speaker information. As discussed earlier, speaker modifies the speech production system to acknowledge the information about the adverse environmental factor. Under stress, both the phonetic and the speaker specific attributes of the speech are affected. Consequently, an ASR system that is trained using the neutral speech and tested on the stressed speech, exhibits a severe degradation in the recognition performance. Therefore, transforming the stressed speech through a subspace projection matrix capturing the principal dimensions of the acoustic variations represented by the neutral speech would help in improving the recognition performance. Furthermore, reducing the rank of the projection matrix will, in turn, reduce the mismatch in the variances resulting from the stress. From the reported works, it is also observed that, for the stressed speech, the high frequency region is affected more in comparison to the low frequency region [19, 41]. Chen [79] has shown that, the cepstral mean values have an exponential spectral tilt under stress. In other work [80], the low-rank features of emotion-specific component, which are derived using the hidden factor analysis (FA) and the mixture of factor analysis (MFA) are explored for speech emotion recognition. In the recently reported works [81, 82] on children's mismatched ASR system, the low-rank subspace projection has been explored and was reported to be very effective for reducing the variance mismatch between the training and the test environments.

In addition to the speech signal, the subspace projection techniques are successfully employed for investigating the stress information using the electroencephalogram (EEG) signal and the facial expression [39, 83–85]. In the work reported in [83], the canonical correlation analysis (CCA) of EEG signal and the privileged information of stimulus video was exploited for emotion recognition. The canonical correlation analysis (CCA) between the EEG signal and the video features were exploited for the creation of new space. This new subspace has resulted in the improved emotion recognition performances by implicit integration of privileged information in the training phase. In the research area of image processing, the subspace projection was accomplished over each pixel vector onto a subspace, which is orthogonal to the undesired signatures [39]. The proposed orthogonal projection was noted to be very effective for the classification of hyper-spectral images. In another work [85], the

orthogonal subspace projection (OSP) technique is generalized and referred to as the generalized OSP (GOSP). The GOSP is used to create the additional multi-spectral bands in unsupervised manner for the implementation of effective processing of hyper-spectral images. The proposed GOSP method was reported to be very efficient for the classification of hyper-spectral images. In the work reported in [84], the restoration of images, which carry the geometric characteristics has been successfully accomplished by orthogonal projection-based linear model. These studies demonstrate the significance of subspace projection approaches for investigating the modification in the characteristics of EEG as well as image features under noisy stressful conditions.

1.4 Speaker Variability Under Stress Condition

As discussed in the Section 1.1, the speaker changes the movement of anatomical structures of the speech production system under stress condition. Speakers with age and gender variations under the similar stress condition produce the same speech utterances with varying parameters of stress. The speaker variability appears more pronouncedly in the stressed speech utterances. The speaker independent (SI) continuous speech recognition system trained on the neutral speech data, when tested using the stressed speech exhibits degraded performances [24, 40, 41, 86]. The inter- and the intra-speaker variabilities causes high overlap between the speech units of neutral and stressed speech and it leads to a large variance mismatch between the training and the test environments. The adaptation of feature-space and model-space onto another subspace, which constitutes the speaker normalized properties would significantly improve the robustness of stress normalization method. In the following, we have discussed the various approaches used in literature to address the speaker variability in the production of stressed speech.

The literature reviews show that, the variation in the spectral-specific and the speaker-specific correlates were found to be very effective to attain the affective computing of human-computer interaction (HCI) systems in case of users under stress condition [24, 58, 59]. In the work reported in [24], the communication between the human and the Thinking Robot (T-ROT) were investigated by developing a speaker-independent system in context of commercial applications. To develop a robust human-robot interaction (HRI) system against the emotional users, they analyzed the spectral properties. The ratio of the geometric mean to the arithmetic mean of the power spectrum, called as the spectral flatness measure (SFM) and the spectral center, which represent the average fre-

quency weighted by acoustic power were studied. In that work, SFM parameters were exploited to capture the spike-like characteristics to differentiate between the voiced and the unvoiced speech. Whereas, the spectral center was used to distinguish between the plosives and the fricatives sounds. Experimental evaluation of the studies shows that, the ratio of SFM to spectral center was an effective index to precisely classify the emotion classes with invariant speaker's information. In addition to this, it was also observed that, the variation in the emotional speech significantly appeared due to the speaker-specific variabilities under the emotional condition. Park et al. [58] have proposed non-overlapped features for speech emotion recognition in context of human-robot interaction. Similar acoustic correlates between the emotions and the speaker-specific variabilities were discarded to derive the non-overlapped features. It was reported that, the inclusion of non-overlapped features in modelling has resulted in the improved accuracy for the classification of boredom, anger, happy, neutral and sadness cases, when compared to the conventional method. In the work reported in [61], significant degree of variation was reported for the pitch and the energy parameters in the production of speech by the female and the male speakers under the lombard effect. Sungrack et al. [33] have proposed the GMM-based discriminant function for the three different speaker-independent emotion classification tasks namely: the acted emotion classification, the natural emotion classification and the cross corpus emotion classification. In that work, the GMM corresponding to each emotion class was trained, in which the margin is scaled by the Hamming loss function for the precise estimation of GMM parameters. All the GMM parameters were estimated using MFCC, pitch, log energy, zero crossing rate, corresponding delta and acceleration coefficients onto the six learning frameworks namely: maximum likelihood (ML), maximum mutual information (MMI), margin scaling using the Hamming loss function (MSH), margin scaling using the linear (MSN), margin scaling using the log (MSL) and margin scaling using the exponential (MSE) function. The authors have observed that, the learning of GMM parameters by employing the MSH and the MSN learning frameworks provide better accuracy for the speech emotion classification and also found that, the resulting GMMs were constituted the speaker invariant properties compared to the parameters of GMMs developed in the ML and the MMI learning frameworks. In the work reported in [81, 82, 87–91], the acoustic mismatch resulting from the speakers with age differences were investigated. In these studies, the acoustic mismatch resulting from the anatomical and the physiological changes during a child's growth causes the greater range of deviation with different means and variances for the pitch, the formant

frequencies, the speaking rate, the average phone duration, the glottal parameters, the grammar and the pronunciation, when compared to the adult speakers have been addressed . The differences in the pitch values lead to a significantly degraded recognition performance. Moreover, it was also observed that, the effect of the pitch-dependent distortions are more pronounced in the lower frequency regions. The vocal-tract length normalization (VTLN) and the low-rank subspace projections were introduced for explicitly normalizing the effect of age variation on the acoustic correlates and found to be very effective in reducing the acoustic mismatch between the speech produced by child and adult speakers. In a recent work [92], authors have been introduced a novel front-end speech parameterization technique employing the adaptive-cepstral truncation prior to the estimation of spectral moments. Experimental evaluations of the studies show that, the proposed acoustic features are very effective in enhancing the pitch-robustness of the existing spectral moment timefrequency distribution augmented by low-order cepstral (SMAC) features. These works show that, the variability due to speakers increases the variance mismatch between the different speech units of neutral and stressed speech. Therefore, the normalization of speaker-specific attributes would significantly help in improving the effectiveness of stress normalization techniques.

1.5 Scope for the Present Investigation

Analysis of stressed speech is important for the development of effective stress normalization technique. The stress normalization method should be robust and computationally efficient for retaining the intelligibility of the stressed speech, respectively. Researchers involved in the area of stressed speech have investigated various methodologies and research issues for an effective processing of stressed speech. Though several works are reported in literature on tasks involving the stress classification and recognition, there is little effort in the direction of stress normalization. Most of the works in the area of robust processing of speech focus on speech under noisy condition. Numerous real life applications involve their interaction with machines for the users under stress condition. Speech produced under stress condition comprises the modified characteristics, when compared to the speech produced under neutral condition. Consequently, the real applications exhibit degraded performances against the users under stress condition. Above facts suggest that, a detailed investigation is required for stress normalization. It is necessary to quantify the acoustic mismatch between the neutral and the stressed speech.

Stressed speech carries vital information about the environmental condition in addition to the acoustically rich phonetic information, which is referred to as the speech information. Separation and detection of stress information may be affected by quasi-stationary speech morphologies, speaker variability and various noises. An effective stress normalization method consists of extraction of features consisting of speech information and precise tuning of model parameters for the robust representation of stressed speech in the feature- and the model-space, respectively. It would be interesting to find out the methodology, which can effectively estimate the robust features and model parameters under stress condition without increasing the computational complexity. There are several signal processing and pattern matching techniques that are widely used for the analysis of stressed speech. The subspace projection approaches motivate the various fields of researchers for the investigation on the subject by creating the subspace corresponding to that subject with fascinating realization properties, effective algorithms and uncomplicated computational characteristics. The analysis of stressed speech using the subspace projection-based method would help in studying the changes in the properties of stressed speech by developing the individual subspaces for speech- and stress-specific attributes. The creation of individual subspaces may help in investigating the relationship between the speech and the stress information and it would lead to the informative direction for normalizing the stress information. There are scopes to exploit the subspace projection methods for the development of effective stress normalization techniques.

The effectiveness of subspace projection depends on the estimation of projection matrix. The subspace projection matrix, whose columns can span the speech-specific attributes can help in reducing the acoustic mismatch between the neutral and the stressed speech signals. Speech information is present in both the neutral and the stressed speech. Therefore, the estimation of projection matrix using the neutral speech utterances would help in retaining the speech information as well as in reducing the mismatch in the variances resulting from the stress. The estimation of low-rank subspace projection matrix provides the effective approach to investigate the properties of stressed speech onto the various bandwidth of frequencies. The subspace projection can be further exploit in sparse domain, where the subspace projection is derived over the over-complete subspace projection matrix called as the dictionary and it is generally referred to as the sparse representation technique. The sparse representation transforms the observed signals linearly using the predefined atoms of dictionary with sparsity. There are scopes to explore the various techniques for the estimation of effective

subspace projection matrix and the mechanism of subspace projection to reduce the aforementioned acoustic mismatch between the neutral and the stressed speech.

In literature, numerous works have investigated the changes in the acoustic parameters of speech production system, which reflects the emotional condition of speakers [5, 60, 62, 64, 93]. The investigation on the modification in the vocal-tract system and the excitation source in time and frequency domains were reported to be very effective for classifying the stress classes [19–22]. The changes in the vocal-tract system by exploring the non-linear air flow patterns, which is modeled using the Teager energy operator (TEO) operator was found to be very effective for the analysis of speech produced under adverse stressful condition [44, 45]. Moreover, significant changes were noted in the source parameters of speech production system under stress condition [25, 26, 68]. The analysis of the precision of articulation and the waveform regularity of successive glottal pulses under stress condition provides the meaningful information about stress environmental factors. There are scopes for the investigation on the vocal-tract system for reducing the acoustic mismatch between the different speech units of neutral and stressed speech. The analysis of vocal-tract system by exploring the subspace projection-based methods can be useful towards the investigation on the changes in the characteristics of stressed speech resulting from the stress.

Furthermore, several researchers have demonstrated that, the speaker-specific variability increases the overlap between the different speech units. The variance mismatch appears more pronouncedly, when speakers are under stress condition. This degrades the performances of various practical applications employing the ASR system trained on neutral speech and tested using stressed speech due to the differences in the training and the test environments. The adaptation of feature- and model-space onto the another subspace, which comprises the speaker normalized attributes would significantly improve the performances of those applications. The speaker dependent speech recognition system is found to be very effective with improved recognition performance for the unknown test speakers [94–99]. The learning of a subspace that consists of phonetic information along with the suppressed speaker variability can help in reducing the variance mismatch between the neutral and the stressed speech. There are scopes to increase the robustness of stress normalizing techniques by suppressing the speaker variability.

The work presented in this thesis is broadly divided into three categories. First, the linear and the non-linear characteristics between the speech- and the stress-specific attributes have been studied

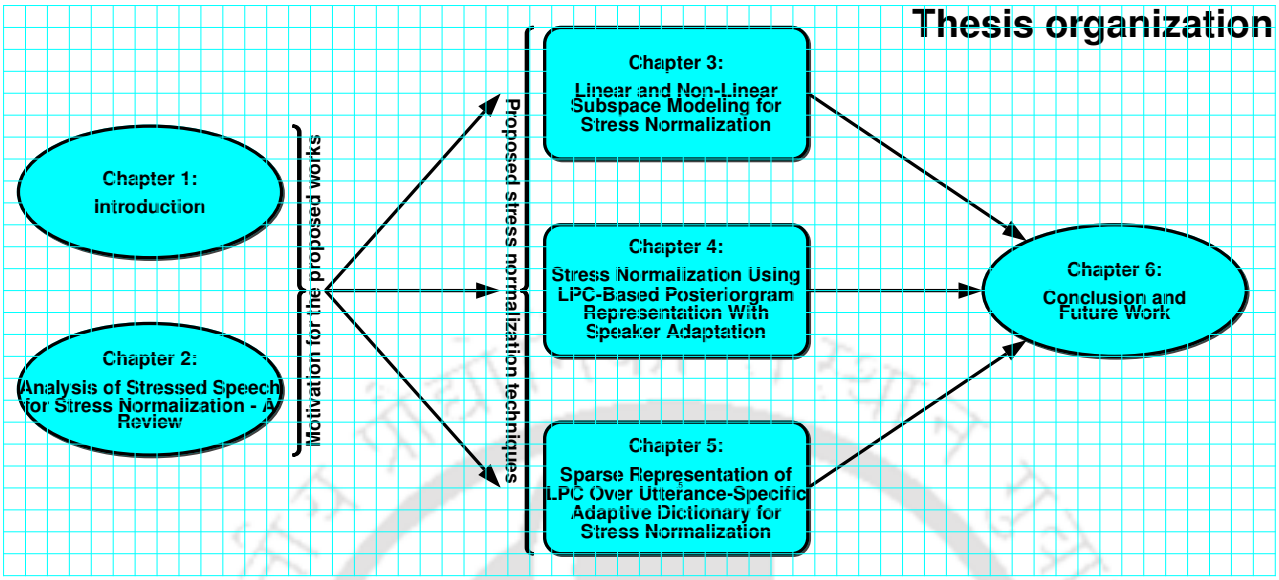


Figure 1.2: Flow chart of organization of proposed works in the thesis.

by exploring the orthogonal projection and the non-linear transformation techniques. Second, the modifications in the characteristics of vocal-tract system under stress condition are investigated by exploiting the subspace projection onto the Gaussian-subspace using the posterior probability information. In the third approach, stressed speech is studied in sparse domain. The acoustic mismatch between the neutral and the stressed speech is reduced by exploring the sparse representation over the utterance-specific adaptive dictionary. Moreover, the effectiveness of all the proposed stress normalization techniques are further improved by reducing the speaker variability. The subspace projection over the low-rank projection matrix has been exploited to investigate the effect of stress in different frequency bands.

1.6 Organization of the Thesis

This Chapter introduced the stress normalization and its processing in acoustic-space, feature-space and model-space from the published literatures. Based on the interpretation described in the previous Sections, in this dissertation, the novel speech subspace modelling with speaker adaptation methods are proposed for the development of effective stress normalization techniques. Figure 1.2 shows the flow chart for the organization of proposed works presented in this dissertation. In the following a brief discussion on the organization of thesis is described.

In **Chapter 2**, a related review of stress normalization are discussed. The acoustic mismatch between the neutral and the stressed speech is studied by analyzing acoustic parameters and three types of front-end speech parametrization techniques. Stress information is further quantified using the stressed speech recognition over the various features and modelling techniques. Linear and non-linear subspace modelling approaches for normalizing the stress information are discussed in **Chapter 3**. The low-rank subspace projection is also introduced to analyze the stressed speech in different frequency bandwidths. In **Chapter 4**, the posteriorgram representation with respect to the LPC is proposed for stress normalization. In this Chapter, the robustness of proposed stress normalization method is further increased by learning a lower dimensional speech subspace comprising the suppressed speaker variability. In **Chapter 5**, sparse representation of LPC over utterance-specific adaptive dictionary is proposed for normalizing the stress-specific attributes. Additionally, a sparse representation of LPC over K-SVD algorithm-based invariable size global dictionary is also explored in this Chapter. Finally, in **Chapter 6**, conclusions are drawn by summarizing the contributions of the dissertation and mentioning some directions for further investigations in these areas.

2

Analysis of Stressed Speech for Stress Normalization - A Review

Contents

The key motivation for researchers is the development of speech parameterization and modelling techniques without degrading the quality of speech signals. The parameterization of speech signal is generally referred to as the feature extraction and it represents the speech samples with few parameters. These parameters, also called as the features are the compressed representation of the acoustic properties manifested in the speech signal and they reflect the characteristics of speech production system [1–3]. The spectral and the cepstral analysis are considered as the state-of-the-art for researchers involved in the area of speech processing to extract the effective features. The Mel-frequency cepstral coefficient (MFCC) [100], the linear prediction cepstral coefficient (LPCC) [1, 101–105], the Teager energy operator (TEO) autocorrelation envelope area (TEO-CB-Auto-Env) features [44, 45], the perceptual linear prediction (PLP) features [106–108], the filter-bank energy [1,19,109–111], the one-sided autocorrelation linear prediction coefficient (OSALPC) [19], the Log-Frequency Power Coefficient (LFPC) [22], the sinusoidal features [41], the derivative of harmonics [26], the spectral envelope-based features [21] are extensively used features for the stressed speech. These features have the significant characteristics of vocal-tract system and excitation source for the production of speech under stress condition. In model-space, the modelling techniques, which are commonly used for the robust and the precise representation of stressed speech features include the vector quantization (VQ) [1], the Gaussian mixture model (GMM) [112] and the hidden Markov model (HMM) [1]. The subspace Gaussian mixture model (SGMM) [52–54] is effectively used to represent the speech patterns in low data scenario. In recent years, the deep neural network (DNN)-based acoustic modelling technique has been widely used because the DNN comprises of properties for the estimation of precisely tuned class probability [47, 48, 113–117]. The fine tuned class probability is successfully used as the posterior probability to model the emission of speech features over each state of HMM and SGMM systems.

As discussed in Chapter 1, the speech signal constitutes the quasi-stationary characteristics and comprises the vital information about phonetic, linguistic, speaker and environmental properties [1–3]. Under stress, the speaker modifies the speech production system to acknowledge about the stress environmental factors and to retain the intelligibility of speech signals. Consequently, the speech produced under stress condition carries the diversified properties compared to the speech produced under neutral condition [4–8]. The various practical applications employing the automatic speech recogntion (ASR) system developed on neutral speech when tested using stressed speech exhibit

degraded performances due to the variance mismatch between the training and the test environments. The normalization of stress-specific attributes also called as the stress normalization would significantly improve the performances of those tasks for the user under stress condition. The stress normalization is increasingly attractive courtesy from a broader range of researchers involved in the area of stressed speech due to it potentiality of robust and precise representation of speech units produced under stress condition. This chapter addresses the literature reviews on the speech parameterization and the modelling techniques used for evaluation of the stress normalization techniques.

The remainder of this Chapter is organized as follows: The description on the databases used for the analysis of stressed speech is presented in Section 2.1. In Section 2.2, the various speech parameterization approaches are discussed. Section 2.3 contains the detail study on the different modelling techniques used for the robust and the precise representation of stressed speech patterns. The motivation for the present research works are summarized in Section 2.4.

2.1 Stressed Speech Databases

The foremost requirement for the investigation on the characteristics of stressed speech is the collection of speech utterances from speakers under stress conditions or the development of stressed speech database. In literature, various studies in the field of stressed speech have motivated the collection of speech data coloured with various stress conditions. Moreover, the literature review manifested that, the visual, the audio-visual and the electroencephalogram (EEG) signals also convey the essential information about the intensity and the class of stress. In the following, we present the detail description on the databases used in this dissertation to measure the effectiveness of various proposed stress normalization methods. This is followed by the brief discussion on the state-of-the-art of databases employed for the processing of stressed speech.

In this thesis, two databases have been employed to validate all the proposed stress normalization techniques namely: the Speech Under Simulated Stress Condition (SUSSC) database [118] and the Database of German Emotional Speech (Emo-DB) [119]. The SUSSC database was developed by us in angry, sad, lombard, happy as well as neutral conditions. The speech utterances are recorded in Indian language, Hindi. A total of 15 non-professional speakers (5 females and 10 males) ages from 25 to 40 years old were employed for the recording of speech corpus at semi-anechoic studio of Electro Medical and Speech Technology (EMST) Laboratory, Department of Electronics and Electrical

Engineering, Indian Institute of technology (IIT) Guwahati, India. All the speech utterances were digitized at sampling frequency of 16 kHz, later they were down-sampled to 8 kHz with 16 bits/sample resolution. The vocabulary of the database constitutes 33 isolated words, 28 short phrases, 29 long phrases and 3 passages. This database has been validated by finding the level of stress present in the speech utterances by exploring the tasks involving the human stressed speech processing and the automatic stressed speech processing. The experimental evaluations presented show that, in average, 59.44% of accuracy has been obtained in classifying correctly the stress classes in human stress classification task. The performances of automatic stress classification with the average value of accuracy of 57.66% and 54.53% have been achieved over the VQ and the HMM as classifiers, respectively. The Emo-DB database was collected in neutral and six stress conditions namely: anger, fear, joy, sadness, disgust and boredom, respectively. The speech utterances were recorded by the 10 professional speakers (5 females and 5 males). The speech corpus were extracted from the 10 German utterances, which include the 5 short and the 5 longer sentences. All the speech data were digitized with a sampling frequency of 48 kHz and after that they were down-sampled to 16 kHz. The accuracy of 96.90%, 88.20%, 87.30%, 86.20%, 83.70%, 80.70% and 79.60% were noted for the recognition of anger, neutral, fear, boredom, joy, sadness and disgust speech, respectively.

The development of audio database comprising speech utterances is extensively used database for analyzing the stressed speech. The Speech Under Simulated and Actual Stress (SUSAS) database developed by Hansen et al. [120] is one of the foremost database, which has been widely utilized for investigating the alteration in the characteristics of stressed speech. The speech corpus were collected in the five major categories of tasks namely: the talking styles, the single tracking tasks, the dual tracking tasks, the actual speech under stress and the psychiatric analysis, respectively. The speech utterances were recorded from the 32 speakers in which 19 speakers are female and 23 speakers are male with ages varying from 26 to 76 years old. A total of 16000 utterances were generated and comprises 35 words vocabulary. All the speech data were digitized with sampling frequency of 8000 Hz and 16 bits/sample resolution. In average, the accuracy of 52.83% has been obtained by the listeners in perceiving neutral, angry, loud and slow utterances [121]. The Speech Under Simulated Emotion (SUSE) database was developed by us in two languages namely: Telugu and English [41]. Thirty native Telugu speakers participated in recording the speech utterances under anger, compassion and happiness conditions. A total of 150 utterances were generated in

each stress class. The accuracy of 80% and 70% have been noted for the task involving the human stress classification using the database recorded in Telugu and English, respectively. The other commonly used audio database is the AIBO (a cross-linguistic emotional speech) database, which constitutes the speech corpus in German and English languages. The speech corpus were collected from children's speech, emotional speech, human-robot communication, cross-linguistics and read vs. spontaneous speech [122]. The database in German language was collected from 51 children (30 females and 21 males) aged between 10 to 13 years old. In English language, 30 children's ages varying from 4 to 14 years old had employed for the recording of speech corpus. The speech utterances were recorded in joyful, surprised, emphatic, helpless, irritated, angry, motherese, bored, reprimanding and rest conditions. Using the German data, the performance of word recognition with a accuracy of 76.7% was obtained over the bi-gram language model.

Apart from the database constituting the speech signals, in literature, numerous works have manifested the importance of audio-visual and EEG signals for the recognition and the classification of stress classes. The eNTERFACE'05 audio-visual emotion database comprises the emotionally coloured audio-visual data is intended for the purpose of audio, video and audio-visual emotion recognition and classification tasks [123]. A total of 1166 audio-visual data were collected from 42 speakers (23% female and 77% male) under neutral as well as six emotion conditions namely: anger, disgust, fear, happiness, sadness and surprise, respectively. The video sequences were compressed using the pixel aspect ratio of 1.067 with 25 frames per second. The audio segments were sampled with sampling frequency of 48 kHz and 16 bits/sample resolution. Among 42 speakers, 25 speakers had correctly recognize the emotion classes. The Sustained Emotionally colored Machine-human Interaction using Nonverbal Expression (SEMAINE) database provides the resources for the building of sensitive artificial listener (SAL) for the development of human-computer interaction (HCI) system [57]. This database comprises the audio-visual data recorded from the emotionally coloured speakers. A total of 150 speakers of age range 22 to 60 years old were employed for recording. The raw video data generated by the visual sensor are much larger in comparison to the audio signal. The video data were compressed at a temporal resolution of 789 $\times$ 580 with 8 bits per pixel using five cameras. The DEAP (Database for Emotion Analysis Using Physiological Signals) database contains the EEG and the peripheral physiological signals in arousal, valence, like/dislike, dominance, and familiarity situations [124]. A total of 32 speakers, aged between 19 and 37 years old were participated in

the data recording. The EEG signals were generated with 32 active AgCl electrodes and compressed using the sampling rate of 512 Hz. This database also consists of frontal face video recorded from 22 of the 32 participants. The F1-score values of 0.583, 0.563 and 0.502 were determined using the EEG signal for the classification of arousal, valence, like/dislike conditions, respectively.

2.2 Features for Stressed Speech

The effectiveness of stress normalization techniques depend on the robust and the precise representation of speech samples. The extraction of informative and illuminating features for the neutral and the stressed speech leads to an effective approach for investigating the acoustic mismatch between them resulting from the stress. The following Subsection describes the literature review of the commonly used features. This is followed by the evaluation of these features for the stressed speech using the statistical measure by exploring the variance analysis.

2.2.1 Feature Extraction

The spectral envelop and the supra-segmental features consist of essential characteristics of vocal-tract system and excitation source, respectively. These features help in investigating the changes in the properties of speech production system under stress condition. In the following paragraphs, we have summarized a brief discussion on the Mel-frequency cepstral coefficient (MFCC) [100], the Teager energy operator (TEO) autocorrelation envelope area (TEO-CB-Auto-Env) [44, 45], the linear prediction coefficient (LPC) [1, 101–105], the perceptual linear prediction (PLP) coefficient [106–108] and the filter-bank energy [1, 19, 109–111], which constitute the significant characteristics of vocal-tract system and excitation source.

Mel-Frequency Cepstral Coefficient (MFCC): The Mel-frequency cepstral coefficients (MFCCs) are the most commonly used features in the area of speech processing. The Mel-frequency cepstrum representation was motivated by the computation of features having an ability to convey the perceptually relevant information. MFCCs capture the variation in the spectral properties of speech and they are determined by processing the speech signal through the triangular bandpass filter-bank

comprising Q filters as follows,

$$\text{MFCC}(i) = \sum_{q=1}^{Q} Y_q \cos\left[i(q - \frac{1}{2})\frac{\pi}{20}\right], \qquad i = 1, 2, \cdots, I \tag{2.1}$$

where, MFCC(i) represents the i^{th}, $(1 \leq i \leq I, I$ denotes the total number of cepstrum coefficients) cepstrum coefficient. The log-energy output of the q^{th} filter is represented by Y_q. The bandwidth of each filter is derived on the Mel-scale to model the characteristics of human perception. Psychophysical studies show that, the human auditory system does not perceive the physical frequency of tone in linear scale [125]. It was observed that, the human perception of sound spectrum follows a Mel-scale. A Mel is a unit for the measurement of perceived frequency of a tone. The mapping of physical frequency on the Mel-scale is approximated as,

$$f_{\text{Mel}} = 2595 \log(+f_{\text{Linear}}/700) \tag{2.2}$$

where, f_{Linear} and f_{Mel} represent the physical frequency and the Mel-frequency. The bandwidth of a critical band in Mel-scale varies linearly between 100 Hz to 1000 Hz of physical frequency and then increases logarithmically. It has been found that, the perception of particular frequency by the auditory system signifies the energy in a critical band around that frequency. Davis et al. [100] have studied the detail of cepstral features on the Mel-scale for the large vocabulary syllable-oriented continuous speech recognition system. MFCCs were noted to be very potent in retaining the significant acoustic information in low frequency region and suppressing the insignificant spectral information in the higher frequency bands. The work reported in [126] illustrates that, the combination of static MFCCs and dynamic MFCCs in MFCCs features, which are determined using the first- and the second-order temporal derivatives of MFCCs, respectively, has resulted in very robust and informational for the development of speaker independent isolated word recognition system.

Teager Energy Operator (TEO) Autocorrelation Envelope Area (TEO-CB-Auto-Env): The vocal-tract system carries the essential information about the speech signal [1–3]. In the work reported in [44, 45], the detailed analysis on the modification in the vocal-tract system under stress condition is presented. Authors have found that, the air propagates within the vocal-tract system in a non-linear trend under adverse stressful condition. This non-linearity in the air flow was modeled using the Teager energy operator (TEO) and it reflects the variation in the airflow during the production

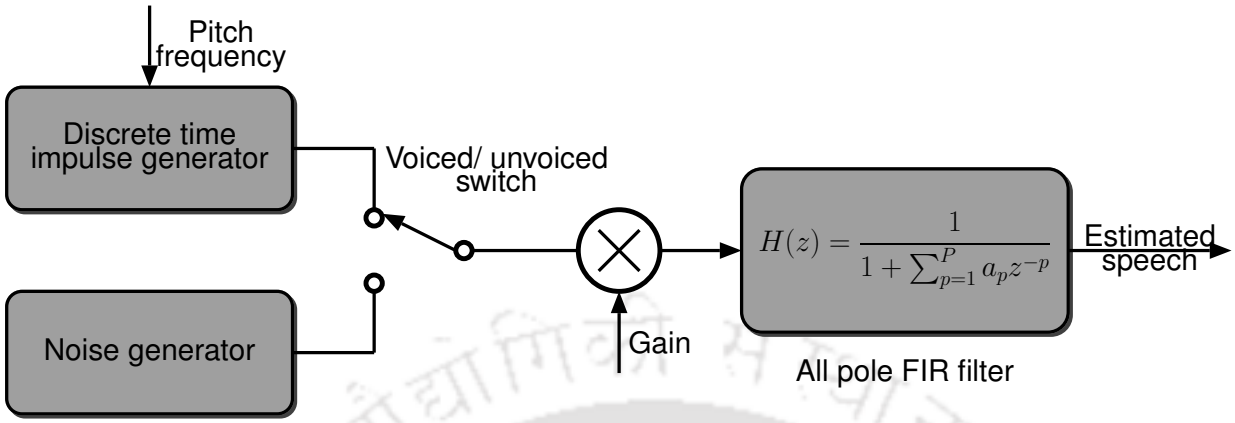


Figure 2.1: Estimation of speech signal production using the linear prediction (LP) analysis.

of speech under stress condition. TEO-CB-Auto-Env features were extracted by filtering the speech signal through the Gabor bandpass filter. The critical band of each filter in filter-bank is determined by the precise partition of audio frequency range on the Mel-scale. After that, the TEO processing is performed over each filtered speech to obtain the TEO profile. The normalized autocorrelation envelope area corresponding to each TEO profile is considered as the TEO-CB-Auto-Env features. The experimental evaluations presented show that, TEO-CB-Auto-Env features capture the text independent information and model the nonlinear airflow, which cause listeners to perceive the information about the stress condition. These characteristics of TEO-CB-Auto-Env features increase both the accuracy and the reliability of tasks involving the classification and the recognition of stress classes.

Linear Prediction Coefficient (LPC): The earliest application of linear prediction (LP) for the analysis of speech is studied by Atal and Hanauer [104, 105, 127]. The linear prediction analysis is a standard source-filter model based approach. In the source-filter model, all-pole finite impulse response (FIR) filter is used to model the slowly varying vocal-tract system. It is hypothesized that, the speech signal is produced by the convolution of fast varying excitation source and slowly varying vocal tract system. The estimation of speech samples using the LP analysis is depicted in the block diagram shown in Figure 2.1. The current sample $s(n)$ of speech signal can be predicted as the linear combination of past P samples and it is computed as follows,

$$\tilde{s}(n) = -\sum_{p=1}^{P} a_p s(n-p) \tag{2.3}$$

$$e(n) = s(n) + \sum_{p=1}^{P} a_p s(n-p) \tag{2.4}$$

The estimated speech sample is represented by $\tilde{s}(n)$. The order of LP analysis is denoted by P. The poles $\{a_p\}_{p=1}^{P}$ of filter are used as the coefficients of linear combination and they are generally referred to as the linear prediction coefficients (LPCs). $e(n)$ represents the residual error. LPCs are estimated my minimizing the instantaneous error between the current sample and the estimated sample. The LPC and the residual error model the significant characteristics of the vocal-tract system and the excitation source, respectively. As discussed earlier, speaker modifies the speech production system under stress condition. Therefore, the LP analysis leads to an effective approach for investigating the changes in the vocal-tract system and the excitation source under stress condition. Moreover, the linear prediction cepstral coefficients (LPCCs), which are derived from the LPC spectrum were reported to be effective for increasing the performances of speech recognition, speaker recognition, speaker verification and speaker identification tasks [101–105]. In addition to these characteristics, the LP analysis is widely used to estimate the pitch frequency from the residual error.

Perceptual Linear Prediction (PLP) Coefficient: The perceptual linear prediction (PLP) analysis is one of the commonly used technique to estimate the human auditory spectrum [107,108]. The PLP technique has exploited the three theory of psychophysics of hearing: the critical-band spectral resolution, the equal-loudness curve and the intensity-loudness power law, respectively. The extraction process of PLP cepstral coefficients is described in the block diagram shown in Figure 2.2. At first, the power spectrum of framed speech signal is determined on the Bark-scale. After that, the auditory warped spectrum is convoluted with the power spectrum of the simulated critical band masking curve and down sampled to almost 1 Bark. In the next step, the power of each frequency bins are mapped according to the power law of Stevens by increasing power by 0.33 in magnitude [125]. The resulting auditory warped line spectrum is processed through the linear prediction analysis to obtain the LPC spectrum. Finally, cepstral coefficients are obtained by a recursion from the LPC spectrum and they are considered as PLP features. Hermansk [107] has observed that, the PLP model with 5^{th} order autoregressive model effectively reduces the speaker variability. These properties of PLP features were reported to be useful for developing the speaker independent ASR system. Moreover, the human auditory system is less insensitive towards the slow and the fast varying frequency components in comparison to the average range of changes in the speech signals [128]. To model these charac-

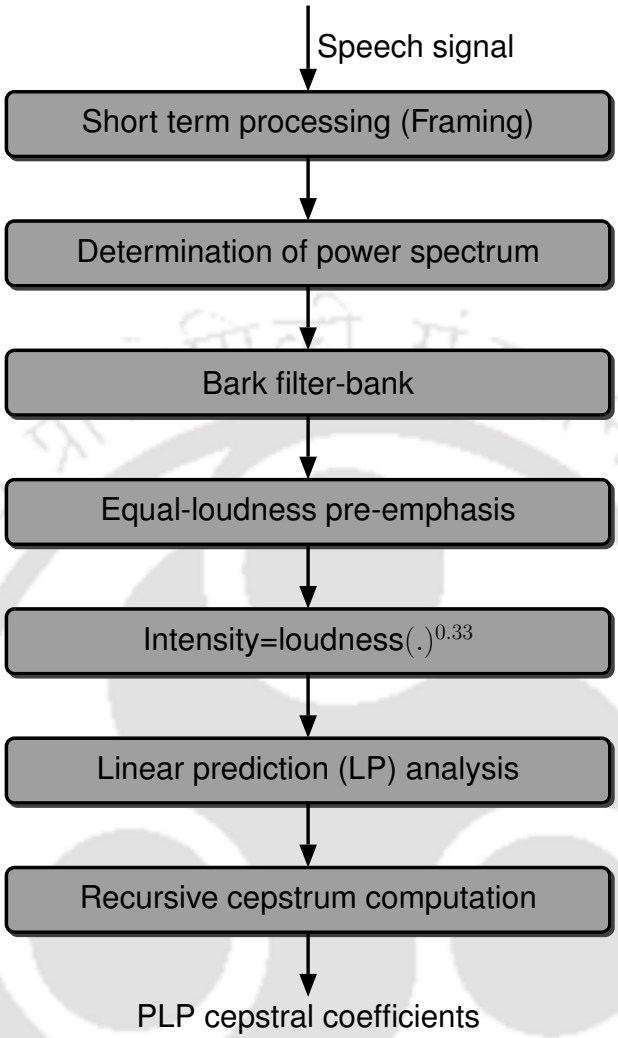


Figure 2.2: Estimation of perceptual linear prediction (PLP) coefficients.

teristics, PLP features are processed through the relative spectral (RASTA) technique and they are called as RASTA-PLP features [106]. The RASTA technique suppresses the slow and the fast varying spectral components using the bandpass filter having a sharp spectral zero at the zero frequency. RASTA-PLP features were noted to be robust in presence of convolutional and additive noises.

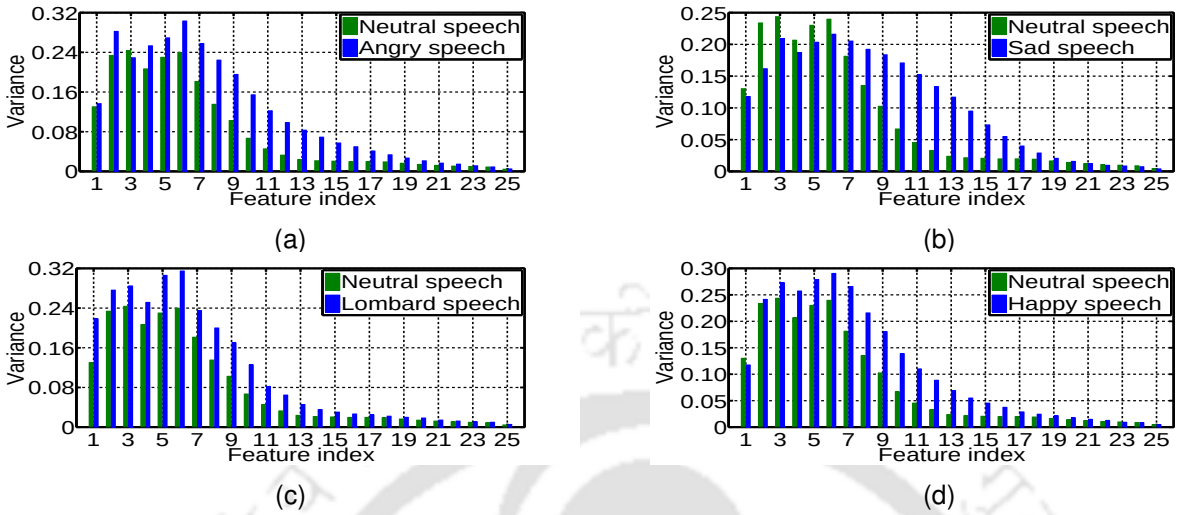
Filter-Bank Energy: The filter-bank analysis has become increasingly popular spectral analysis technique [1, 19, 109–111]. This technique represents the energy at each frequency band of interest in a speech signal using the highly overlapped bandpass filters. These bandpass filters approximate the frequency response of basilar membrane in the cochlea of human auditory system. The output of these filters are considered as the short time spectral envelops consisting of essential information

of speech in different audio frequency bands. One of the conventional method for the computation of filter-bank energy employs the short term Fourier transform (STFT) followed by the low pass filtering technique [1]. In other work [126] , frequency bands are determined on the Mel-scale and energy in each frequency band is weighted and averaged for the precise representation of human hearing system. Furthermore, the combination of filter-bank analysis and linear prediction analysis have been also proposed [107]. In the work reported in [111], the discriminating ability of Mel-frequency cepstral coefficients (MFCCs) is improved by the weighted filter-bank analysis (WFBA). Climent et al. [110] have investigated the designing of filters by combining the frequency filtering transformation and the time filtering transformation to develop the robust speech recognition system.

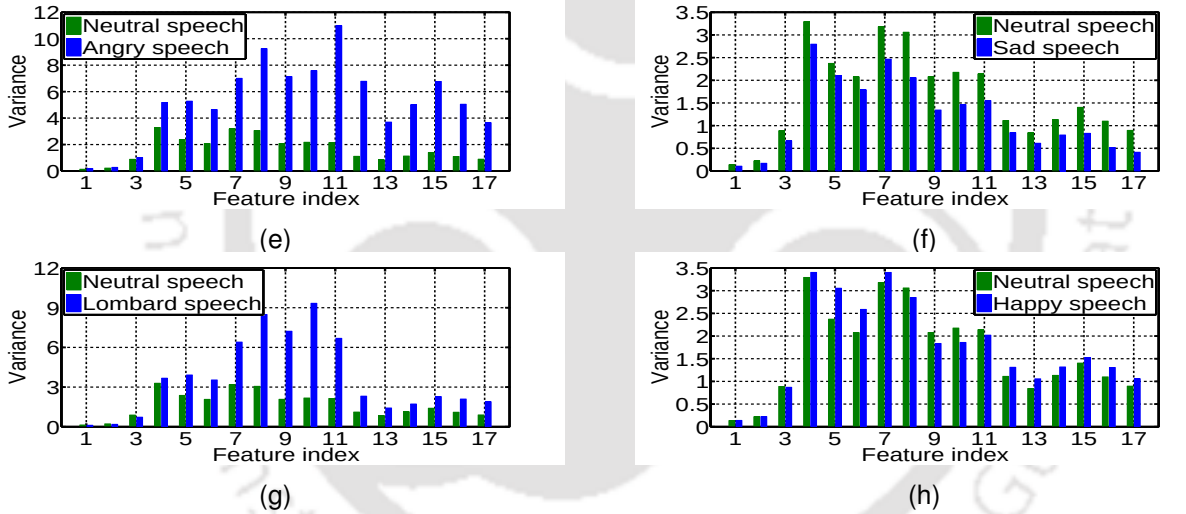
2.2.2 Evaluation of Stressed Speech Features

In this Subsection, the acoustic mismatch between the speech produced under neutral and stress conditions is measured by evaluating the discrepancy between their respective speech features. The discrepancy between the features of neutral and stressed speech is quantified by implementing an experimental analysis using the SUSSC databases reported in [118] as described in Section 2.1. To measure the acoustic dissimilarities, three types of front-end speech parametrization techniques namely: the LPC [1, 101–105], the RASTA-PLP [106] and the MFCC [100] features as discussed in preceding Subsection 2.2.1 have been explored. The statistical measure as the variance analysis has been exploited to quantify the aforementioned discrepancy. Approximately, 2000 frames of speech utterances of two kinds of word, /angoothi/ and /daakghar/ recorded by different speakers are employed for the evaluation of those features. In this work, LPC features constitute 25 LP coefficients and they are extracted using the 25-order LP model. The 17-dimensional RASTA-PLP features are determined by processing the speech through the 21-channel Bark filter-bank. The MFCCs features consisting of 13 cepstral coeffficients ($C1 - C13$) are computed using the 21-channel Mel filter-bank.

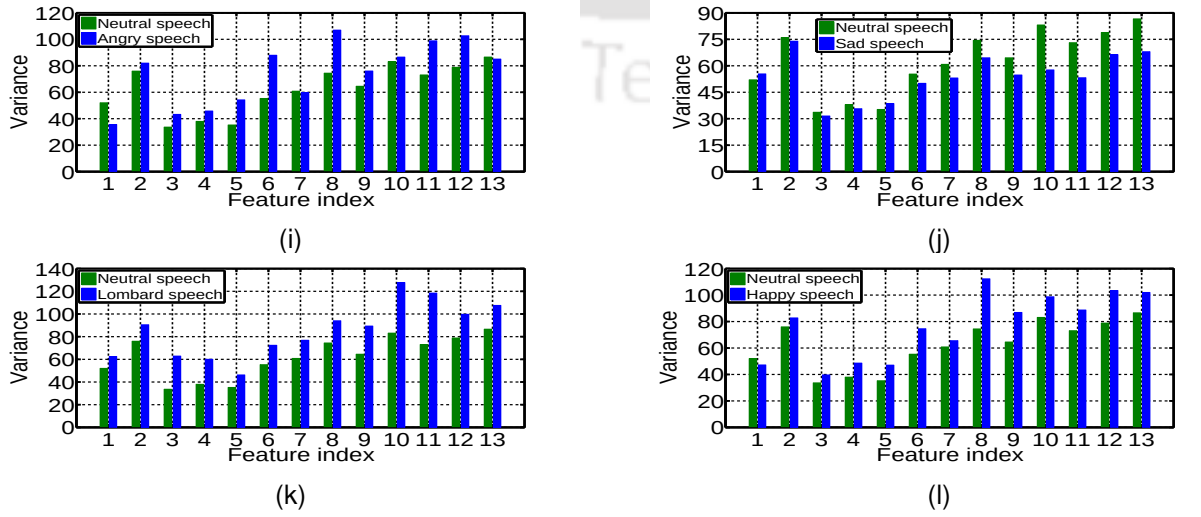
The variance analysis for the LPC, the RASTA-PLP and the MFCC features over both the exploited words are summarized using the bar plots as shown in Figure 2.3 and Figure 2.4, respectively. The significant degree of variance mismatches are reported between the features of neutral and stressed speech for all speech parameterization techniques and stress classes explored in this study. It is to note that, a larger difference of variances are obtained between the LPCs of neutral and speech produced under angry, sad, lombard and happy conditions as shown in Figure 2.3(a)−Figure 2.3(d) and Figure 2.4(d)−Figure 2.4(a), respectively. Figure 2.3(f) and Figure 2.4(f) show the greatly reduced



Variance analysis for the LPC features; (a) angry, (b) sad, (c) lombard and (d) happy

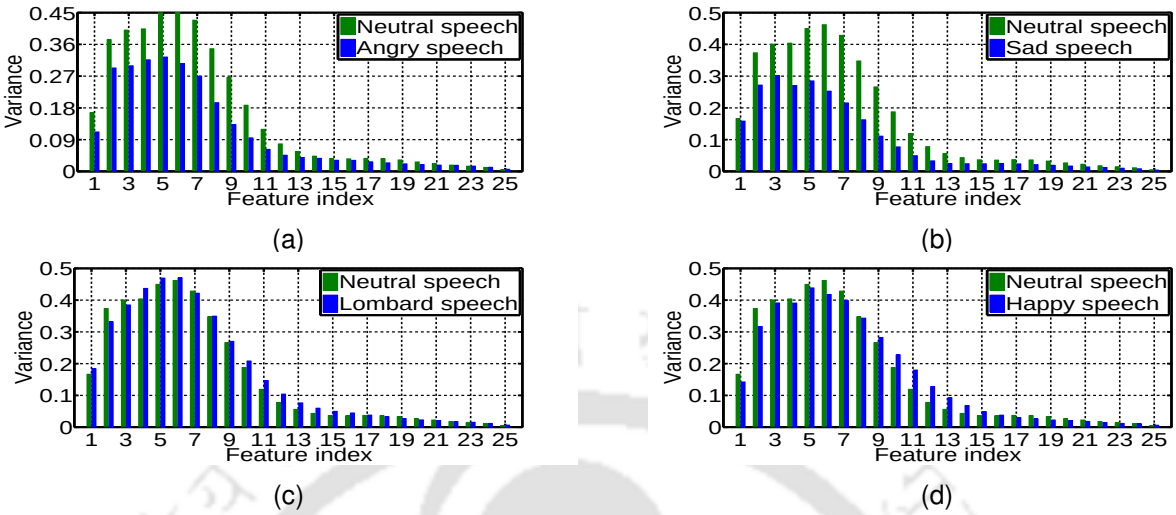


Variance analysis for the RASTA-PLP features; (e) angry, (f) sad, (g) lombard and (h) happy

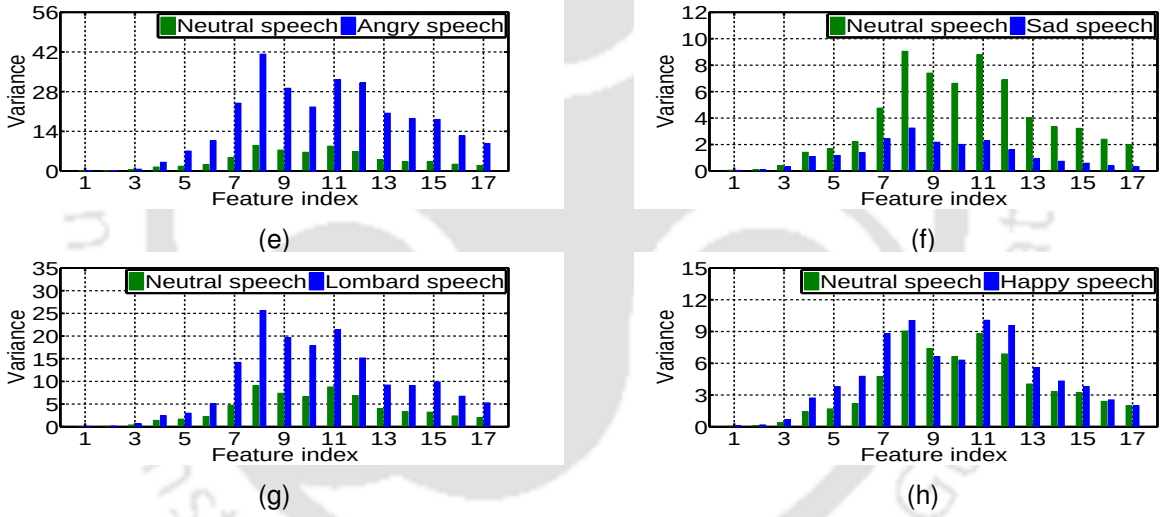


Variance analysis for the MFCC features; (i) angry, (j) sad, (k) lombard and (l) happy

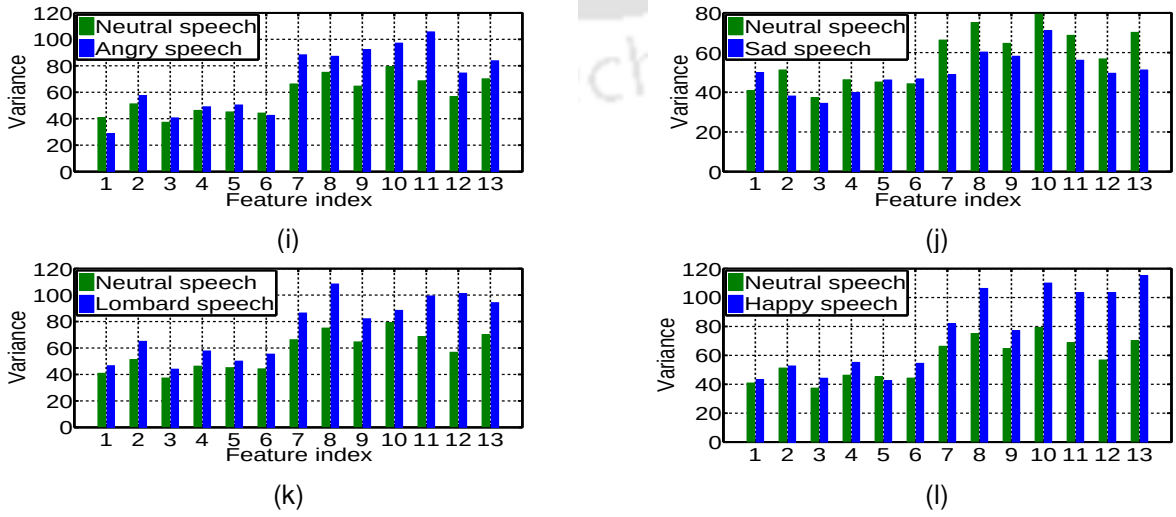
Figure 2.3: Bar plots for the variance interpretations of LPC, RASTA-PLP and MFCC features of neutral and stressed speech utterances of word /angoothi/.



Variance analysis for the LPC features; (a) angry, (b) sad, (c) lombard and (d) happy



Variance analysis for the RASTA-PLP features; (e) angry, (f) sad, (g) lombard and (h) happy



Variance analysis for the MFCC features; (i) angry, (j) sad, (k) lombard and (l) happy

Figure 2.4: Bar plots for the variance interpretations of LPC, RASTA-PLP and MFCC features of neutral and stressed speech utterances of word /daakghar/.

values of variance for RASTA-PLP features of sad speech of both the words compared to the variances computed using the neutral speech utterances, respectively. For angry and lombard speech, variances of RASTA-PLP features increased in a large scale in comparison to those determined in case of using neutral speech utterances as depicted in Figure 2.3(e), Figure 2.3(g), Figure 2.4(e) and Figure 2.4(g), respectively. The similar effects of variance mismatches are observed between the MFCC features of neutral and stressed speech utterances of all the explored stress classes as shown in Figure 2.3(i)−Figure 2.3(l) and Figure 2.4(i)−Figure 2.4(l), respectively. Using MFCC features, the variance metric exhibits the increased values for angry, lombard and happy speech compared to the variances of MFCCs for the neutral speech of both the studied words as depicted in Figure 2.3(i), Figure 2.3(k), Figure 2.3(l), Figure 2.4(i), Figure 2.4(k) and Figure 2.4(l), respectively. Whereas, the degraded values of variances are noted for the MFCC features of sad speech in comparison to the MFCC features of neutral speech of both the tested words as shown in Figure 2.3(j) and Figure 2.4(j), respectively. These differences between the variances for all speech parameterization techniques and stress classes explored in this work acknowledge about changes in the properties of human auditory and speech production system under stress condition. The spectral contents of stressed speech in different frequency bandwidths carry the modified characteristics. These modified characteristics increase the variance mismatch between the spectral envelopes of neutral and stressed speech signals.

2.3 Modeling Approaches for stressed speech

Stress modifies the characteristics of speech and it introduces the acoustic mismatch between the neutral and the stressed speech signals. This phenomenon is also highlighted from the observations of the experimental study carried out in Subsection2.2.2. The robust and the precise optimization of model parameters help in effective representation of stressed speech with less overlapped between the different speech units in model-space. In the following Subsection, we have summarized the literature reviews on the various acoustic modelling approaches. The succeeding Subsection contains the evaluation of these modelling techniques for the stressed speech.

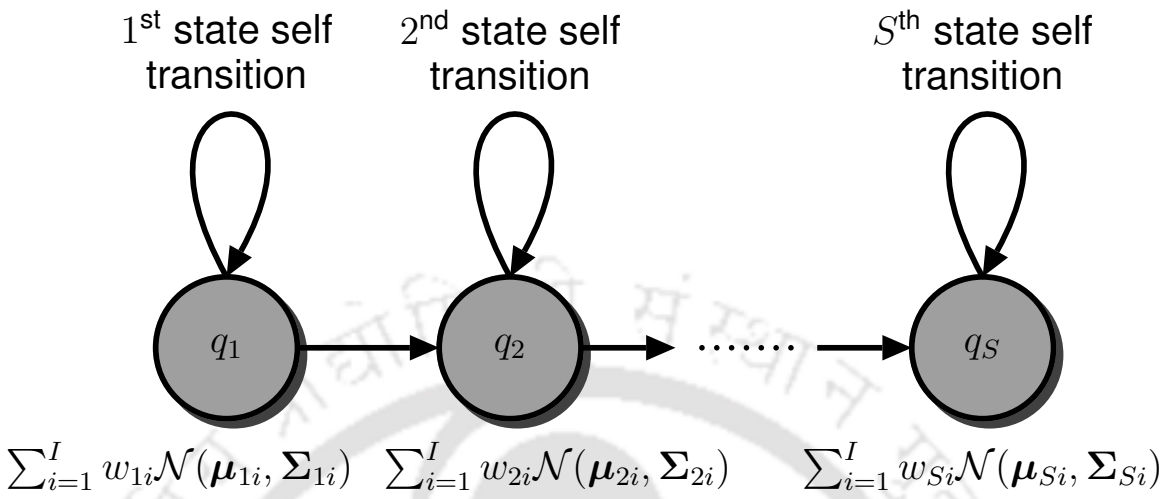


Figure 2.5: The left-to-right continuous density hidden Markov model comprising S states.

2.3.1 Acoustic Modeling Techniques

There are growing research in the field of stressed speech by exploring the robust and the explicit modelling of anatomical and physiological changes during stress. These changes are modeled using the various learning algorithms in supervised, semi-supervised and unsupervised manners to optimized the fine tuned parameters of acoustic model for the robust modelling of stressed speech patterns [27–29]. In the following paragraph, we have presented a brief discussion on the acoustic modelling techniques employing the Gaussian mixture model (GMM) [1], the subspace Gaussian mixture model (SGMM) [52–54] and the deep neural network (DNN) [55, 56].

Gaussian Mixture Model (GMM)-Based Acoustic Modeling: The Gaussian mixture model (GMM)-based acoustic modelling technique employs the multivariate Gaussian mixture model to represent the speech units in each state of hidden Markov model (HMM) [1]. In this thesis, the GMM-based acoustic model is referred to as the GMM-HMM. The GMM-HMM system estimates the stationary interval of quasi-stationary speech using a single state by exploiting the parametric stochastic random process. The structure of left-to-right GMM-HMM system comprising S, $(1 \leq s \leq S)$ states is shown in Figure 2.5. The speech features $\mathbf{x}_t$ corresponding to each frame are modeled as the output of

states using the continuous probability density function $p_s(\mathbf{x}_t)$ as follows,

$$p_s(\mathbf{x}_t) = \sum_{i=1}^{I} w_{si}\mathcal{N}(\mathbf{x}_t, \boldsymbol{\mu}_{si}, \boldsymbol{\Sigma}_{si}) \tag{2.5}$$

$$\sum_{i=1}^{I} w_{si} = 1$$

and

$$\int_{-\infty}^{\infty} p_s(\mathbf{x}_t) d\mathbf{x}_t = 1$$

where, $\mathcal{N}$ denotes the continuous density Gaussian mixture model, which comprises the weighted sum of I, $(1 \leq i \leq I)$ Gaussian densities. The mean $\boldsymbol{\mu}_{si}$, the variance $\boldsymbol{\Sigma}_{si}$ and the weight w_{si} parameters of GMM-HMM system are estimated using the expectation-maximization (EM) algorithm to model the likelihood of the speech features [129, 130]. In last four decades, the GMM-HMM system is successfully employed in the large scale laboratory and the commercial speech recognition system [131–134]. In the area of stressed speech, the GMM-HMM system is one of the extensively used classifier for tasks involving the stressed speech recognition and classification [135–137]. In another work [27], the $N-$channel HMM is developed by allowing the transition between one-dimensional HMMs. The $N-$channel HMM system was reported to be very effective for capturing the divergences in the various acoustic parameters resulting from stress.

Subspace Gaussian Mixture Model (SGMM)-Based Acoustic Modeling: The subspace Gaussian mixture model (SGMM) approach consists of hidden Markov model (HMM) and share a common Gaussian mixture model for all the phonetic states [52–54]. In this thesis, the SGMM-based modelling technique is represented by SGMM-HMM. The mean $\boldsymbol{\mu}_{si}$ and the weight $\mathbf{w}_{si}$ parameters are varied in a full parameter subspace having the Gaussian structure. Each state s, $(1 \leq s \leq S)$ of SGMM-HMM is characterized by the specific vector parameter $\mathbf{h}_s$ and the global shared matrix $\mathbf{H}_s$. This global shared matrix transform the specific vector parameter associated with each state for the estimation of mean and weight parameters. The SGMM-HMM model is represented by the updated mean, weight and likelihood of features $\mathbf{x}_t$ in each state of HMM and they are determined as follows,

$$p(\mathbf{x}_t \mid s) = \sum_{i=1}^{i} w_{si}\mathcal{N}(\mathbf{x}_t \mid \boldsymbol{\mu}_{si}, \boldsymbol{\Sigma}_i) \tag{2.6}$$

$$\boldsymbol{\mu}_{si} = \mathbf{H}_i\mathbf{h}_s \tag{2.7}$$

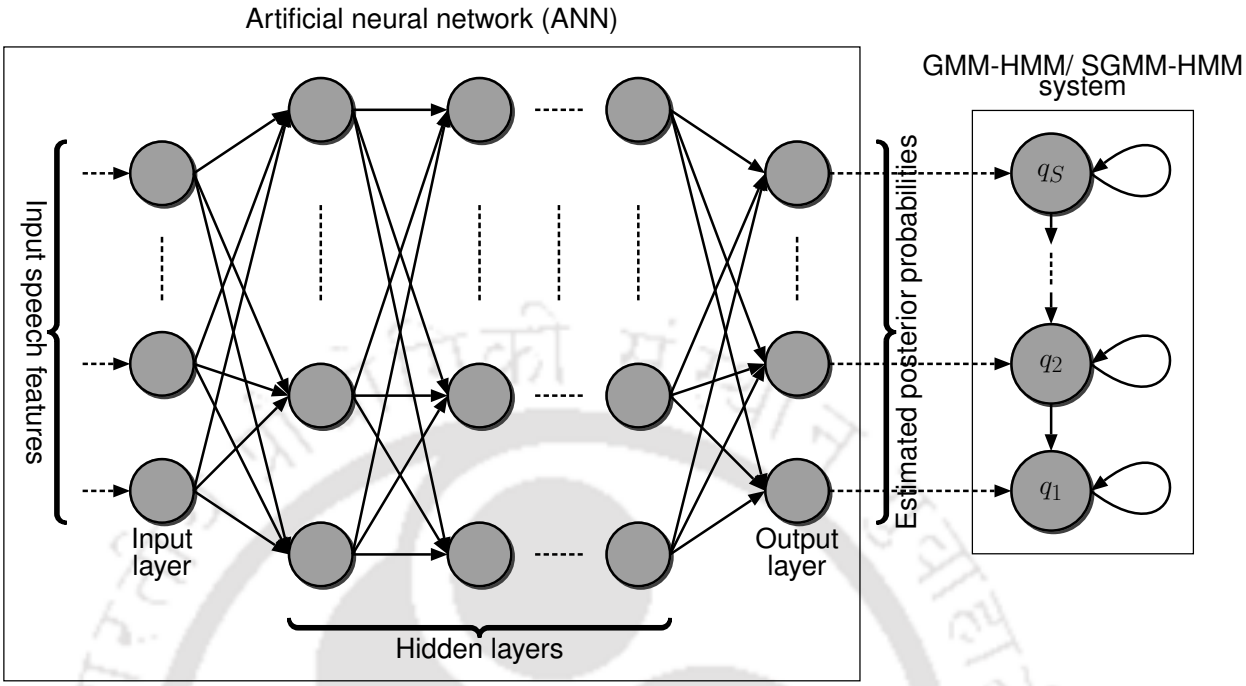


Figure 2.6: The structure of deep neural network (DNN)-based acoustic modelling technique.

$$w_{si} = \frac{\exp(\mathbf{w}_i^T \mathbf{h}_s)}{\sum_{i=1}^{I} \exp(\mathbf{w}_i^T \mathbf{h}_s)} \tag{2.8}$$

the EM technique has been incorporated for the optimization of mean $\boldsymbol{\mu}_{si}$ and weight $\mathbf{w}_{si}$ parameters [129, 130]. The SGMM-HMM modelling technique shares the full parameter Gaussian subspace over each states of HMM, this, in turns, reduces the number of parameters for precise and robust modelling of speech units. The SGMM-HMM system gives the better accuracy particularly with smaller amounts of training data in comparison to the GMM-based acoustic modelling technique.

Deep Neural Netwotk (DNN)-Based Acoustic Modeling: In recent past, the deep neural network (DNN) [55, 56] has been successfully employed in speech recognition. In most of the reported works, the speech recognition systems employing the DNN-based acoustic modelling have been reported to improved performances those based on the GMM and the SGMM systems. The DNN system employs many layers of non-linear hidden units and a very large output layer to model the acoustic variations. Its the posterior probabilities of the context-dependent tied state or the senones are estimated using the DNN. These posterior probabilities are then used in the GMM- and the SGMM-based acoustic modelling technique. Figure 2.6 shows the topology of DNN-based acoustic modelling tech-

nique. An artificial neural network (ANN) with more than one hidden layers between the input and output layers is used to create deep neural network. For the input data vector for an ANN, the activation probability vector for the hidden unit is computed. The activation probability vector is then used as an input to train the next hidden layer. Thus, for any particular layer, the set of weights are actually the non-linear features for the output of the previous layer. After that, a randomly initialized soft-max output layer is then added. Next, all the weights of the network are discriminatively fine tuned by back-propagating the error. This can be done by back-propagating derivatives of the cost function i.e., the cross-entropy between the target probabilities and the output of the soft-max function. The soft-max non-linearity is used to convert the total input at the output unit into a class probability. The senone likelihood is incorporated as the input to produce the posterior probabilities over the states of the GMM-HMM and the SGMM-HMM systems as depicted in Figure 2.6. In this thesis, this modelling technique of incorporation of the senone likelihood of DNN over the states of GMM-HMM and SGMM-HMM systems are denoted by DNN-HMM and DNN-SGMM, respectively. In literature, few studies on DNN-based ASR system are reported in the area of stressed speech [115, 138]. Stress leads to a non-uniformity between the alignment of state sequences of neutral and stressed speech units. This mismatched alignment of state sequences degrades the performances of ASR system trained on the neutral speech, when tested using the stressed speech. The DNN learns the non-linear hierarchy in data space [55]. Consequently, the output of DNN has resulted in a improved estimate of posterior probabilities over the states of HMM [1] and SGMM [52–54] systems. The non-linear characteristic between the input and the output of DNN helps in reducing the mismatches between the alignment of state sequences and it yields the robust and the precise representation of stressed speech patterns.

2.3.2 Evaluation of Acoustic Models

This Subsection addresses the acoustic mismatch between the speech units of neutral and stressed speech by evaluating the performances of acoustic modelling approaches studied in the preceding Subsection 2.3.1. The automatic speech recognition (ASR) systems are developed by exploring the GMM-, the SGMM- and the DNN-based acoustic modelling techniques. To measure the acoustic dissimilarities, the acoustic models are trained using the neutral speech data and decode using the stressed speech utterances. The block diagram shown in Figure 2.7 details the steps involved in the recognition of stressed speech. A word error rate (WER) metric is used for evaluating the perfor-

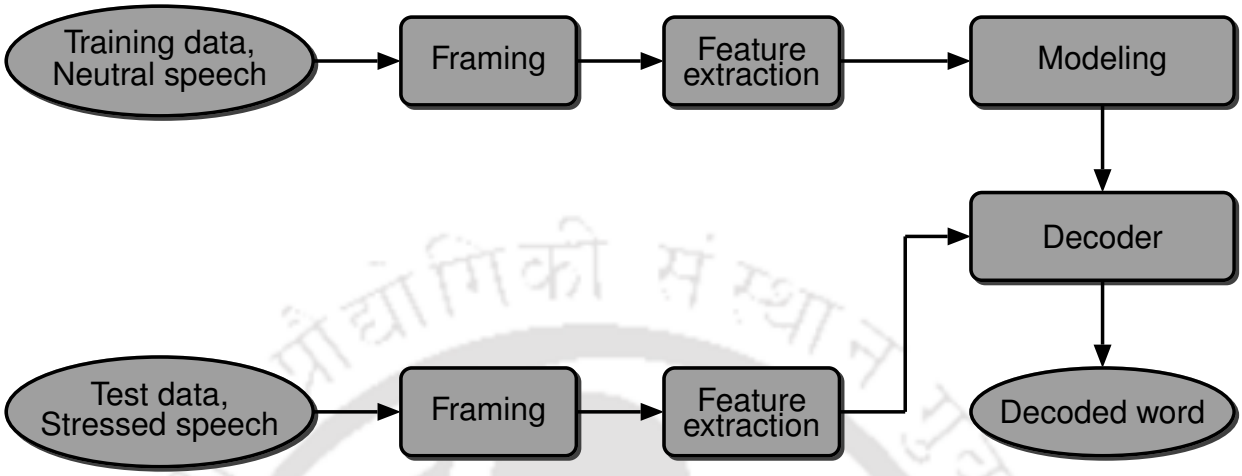


Figure 2.7: The block diagram for the speech recognition system trained on the neutral speech data and tested using the stressed speech utterances.

mance of various ASR systems developed in this work and it is computed as follows,

$$\%\text{WER} = \frac{\text{Sub} + \text{Del} + \text{Ins}}{\text{Total number of words}} \times 100 \tag{2.9}$$

where, Sub, Del and Ins denote the number of substitutions, deletions and insertions, respectively. The following paragraphs describe the experimental setup and the performance evaluation of stressed speech recognition with respect to the explored speech parametrization and modelling techniques.

Experimental Setup: The performance of the stressed speech recognition is evaluated using the databases reported in [118] as discussed in Section 2.1. Fifteen native speakers (5 females and 10 males) have collaborated for the recording of speech utterances under neutral, angry, sad, lombard and happy conditions, respectively. The acoustic model is trained using 2322 utterances of neutral speech. The test data consists of 700, 700, 594, 594 and 700 utterances of neutral, angry, sad, lombard and happy speech, respectively. All speech data used are digitized at a sampling frequency of 8 kHz with 16 bits/sample resolution and interpreted using a Hamming window of length 20 msec, with a frame shift of 10 msec. Four kinds of speech features namely: the 39-dimensional Mel-

frequency cepstral coefficients (MFCCs) [100], the 39-dimensional Teager energy operator (TEO) autocorrelation envelope area (TEO-CB-Auto-Env) features [44, 45], the 17-dimensional perceptual linear prediction PLP features [106–108] and the 21-dimensional filter-bank energy [1, 19, 109–111] have been explored for parameterizing the speech utterances as described in Subsection 2.2.1. First 12-dimensional base MFCCs ($C1 - C12$) are determined using a 21-channel Mel filter-bank. Energy is added as the zeroth coefficient making the base feature dimension equal to 13 $(C0 - C12)$. These features are coupled with their first- and second-order temporal derivatives to yield 39-dimensional features. All the explored speech features are pre-processed by introducing the cepstral mean subtraction followed by the cepstral variance normalization (CMVN) techniques.

In modelling paradigm, we have explored the speaker dependent (SD) ASR systems employing acoustic models based on Gaussian mixture model (GMM) [1], subspace Gaussian mixture model (SGMM) [52–54] as well as deep neural network (DNN) [55, 56]. The acoustic models are developed using the whole word acoustic modelling along with a decision tree-based state tying. In case of using MFCC, PLP, filter-bank energy as the speech features, the GMM-HMM system consists of a 3-states HMM and the output of each state is modeled using the mixture of 8 diagonal covariance Gaussian densities. The SGMM-HMM model has employed 400 Gaussian densities in the universal background model (UBM). Number of leaves and Gaussian densities in the SGMM are selected as 9000 and 7000, respectively. Using TEO-CB-Auto-Env features, the GMM-HMM system comprises 8-states HMM and 8 diagonal covariance Gaussian components are used to model the output of each state. The SGMM-HMM system consists of 64 Gaussian densities in the UBM with 400 and 500 numbers of leaves and Gaussian densities, respectively. The DNN-based ASR systems employing the GMM-HMM and the SGMM-HMM systems, which are referred to as the DNN-HMM and the DNN-SGMM systems are explored by varying the number of hidden layers 1 to 6, respectively. The input and the output layers consist of 300 and 312 neurons, respectively. Each hidden layer has 300 nodes. The output layer uses soft-max function. The DNN is trained using back-propagation algorithm employing cross entropy as the optimization criterion with *tanh* nonlinearity. Initially, the learning rate is set as 0.005 and then decreased upto 0.0005 for 20 epochs. After that, learning rate is kept constant for next 10 epochs for fine tuning. The size of mini-batch is taken as 128. The input to the DNN comprises the spliced features with a context size of 9 frames. All the explored acoustic modelling techniques are developed using the Kaldi toolkit [139].

Table 2.1: The recognition performances for stressed speech (WER in %). The performances are given for the MFCC, the TEO-CB-Auto-Env, the PLP and the filter-bank energy features with respect to the GMM-HMM and the SGMM-HMM systems.

Feature type	Stress class	Acoustic modelling approach	
		GMM-HMM	SGMM-HMM
MFCC	Neutral	0.71	0.14
	Angry	25.71	48.71
	Sad	**7.41**	**24.07**
	Lombard	**6.73**	**26.26**
	Happy	**7.86**	**28.14**
TEO-CB-Auto-Env	Neutral	**0**	**0**
	Angry	41.71	**44.43**
	Sad	31.31	36.36
	Lombard	24.07	28.79
	Happy	28.29	33
PLP	Neutral	0.14	1.29
	Angry	**21.71**	52
	Sad	10.44	36.70
	Lombard	7.07	28.62
	Happy	8.71	31.86
Filter-bank energy	Neutral	10.71	10.29
	Angry	52.43	72.14
	Sad	42.09	67.85
	Lombard	38.05	61.78
	Happy	33.14	57.43

Performance Evaluation: The performance of speech recognition for the stressed speech with respect to the GMM-HMM and the SGMM-HMM systems are summarized in the Table 2.1. The accuracy of speech recognitions are evaluated using the 39-dimensional MFCC, the 39-dimensional TEO-CB-Auto-Env, the 17-dimensional PLP and the 21-dimensional filter-bank energy features. The bold face WER values represent the best speech recognition performances. The WERs depicted in this table show the severe degradation in the performance for the stressed speech case in comparison to that for the neutral speech. This degradation is consistent across all the feature types as well as the acoustic modelling approaches. The GMM-HMM system trained using the PLP features has resulted in the best speech recognition performance for recognizing the angry speech with minimum WER of 21.71% in comparison to the WER values obtained using other speech parameterization and modelling techniques. For the recognition of sad, lombard and happy speech, the GMM-HMM system developed using the MFCC features has leaded to the maximum speech recognition performances with minimum WERs of 7.41%, 6.73% and 7.86%, respectively. Whereas, using SGMM-HMM system developed on the TEO-CB-Auto-Env features, the maximum deterioration in the WER of 44.43% has been noted for the recognition of angry speech as shown in Table 2.1. Once again, the MFCC features of sad, lombard and happy speech provide the minimum WER values of 24.07%, 26.26% and 28.14% over the SGMM-HMM system. Moreover, it has been also noted that, the GMM-HMM system leads to the improved speech recognition performances in comparison to that obtained by employing the SGMM-HMM system over all the explored features and stress classes.

The WERs reported in the Table 2.1 illustrate that, in most cases, the MFCC and the TEO-CB-Auto-Env features have resulted in the improved speech recognition performances for the stressed speech, when compared to those determined using the PLP and the filter-bank energy features. Therefore, the DNN-HMM and the DNN-SGMM acoustic models have been evaluated for the MFCC and the TEO-CB-Auto-Env features. In this work, the effectiveness of the DNN-HMM and the DNN-SGMM has been investigated by varying the number of hidden layers from 1 to 6. The performance of stressed speech recognition using the MFCC and the TEO-CB-Auto-Env features with respect to the DNN-HMM and the DNN-SGMM systems are tabulated in Table 2.2 and Table 2.3, respectively. The performance of speech recognition over the DNN-HMM systems with 1, 6, 3, 4 and 5 number of hidden layers developed on the MFCC features give the best speech recognition performances with the decrement in the WERs of 1.29%, 18.71%, 9.09%, 7.41% and 8.43% for recognizing the neutral,

Table 2.2: The recognition performances for stressed speech (WER in %). The performances are given for the MFCC features with respect to the DNN-HMM and the DNN-SGMM systems.

Number of hidden layer	Stress class	Acoustic modelling technique	
		DNN-HMM	DNN-SGMM
1	Neutral	1.86	5.86
	Angry	22	32.71
	Sad	13.30	28.45
	Lombard	11.62	21.72
	Happy	11.14	19.43
2	Neutral	1.43	2.14
	Angry	19	25.43
	Sad	11.28	19.53
	Lombard	9.26	**14.98**
	Happy	8.57	13.57
3	Neutral	**1.29**	**1.71**
	Angry	18.86	24.29
	Sad	**9.09**	16.33
	Lombard	8.42	16.16
	Happy	8.57	13.71
4	Neutral	1.29	1.71
	Angry	19.43	23.29
	Sad	10.44	14.65
	Lombard	**7.41**	15.99
	Happy	8.43	13.86
5	Neutral	1.71	2.43
	Angry	18.86	**22.43**
	Sad	9.43	**14.31**
	Lombard	9.93	15.15
	Happy	**8.43**	**12.86**
6	Neutral	1.57	2.43
	Angry	**18.71**	23.57
	Sad	9.93	15.49
	Lombard	9.43	15.82
	Happy	9	13.86

Table 2.3: The recognition performances for stressed speech (WER in %). The performances are given for the TEO-CB-Auto-Env features with respect to the DNN-HMM and the DNN-SGMM systems.

Number of hidden layer	Stress class	Acoustic modelling approach	
		DNN-HMM	DNN-SGMM
1	Neutral	1	1.29
	Angry	13.29	13.29
	Sad	14.31	14.81
	Lombard	7.91	9.60
	Happy	9.57	8.86
2	Neutral	0.71	**0.71**
	Angry	12.14	13.57
	Sad	14.48	14.31
	Lombard	7.58	7.91
	Happy	8	7.71
3	Neutral	0.71	0.86
	Angry	**11.29**	12.57
	Sad	12.96	13.13
	Lombard	**7.41**	**7.74**
	Happy	**7.29**	**7.57**
4	Neutral	**0.43**	0.71
	Angry	12	12.86
	Sad	12.12	13.64
	Lombard	7.74	8.08
	Happy	7.57	8.71
5	Neutral	0.57	0.86
	Angry	12.43	**12**
	Sad	**11.95**	**12.12**
	Lombard	7.91	7.74
	Happy	8	8.29
6	Neutral	0.57	0.86
	Angry	12.71	13.14
	Sad	12.12	13.47
	Lombard	8.42	8.75
	Happy	8.29	9.14

angry, sad, lombard and happy speech, respectively, as depicted in Table 2.2. The DNN-SGMM systems with 3, 5, 5, 2 and 5 hidden layers trained using the MFCC features lead to maximum improvement in the speech recognition performances with the decrement in the WERs of 1.71%, 22.43%, 14.31%, 14.98% and 12.86% for the recognition of neutral, angry, sad, lombard and happy speech, respectively. Using TEO-CB-Auto-Env features, the maximum decrement in the WERs of 0.43%, 11.29%, 11.95%, 7.41% and 7.29% are achieved for the recognition of neutral, angry, sad, lombard and happy speech over the DNN-HMM systems with 4, 3, 5, 3 and 3 hidden layers, respectively as shown in Table 2.3. Whereas, DNN-SGMM systems with 2, 5, 5, 3 and 3 hidden layers developed using the TEO-CB-Auto-Env features have resulted in the maximum improvement in the recognition performances with the decrement in the WER values of 0.71%, 12%, 12.12%, 7.74% and 7.57% for recognizing the neutral, angry, sad, lombard and happy speech, respectively. Furthermore, the DNN-HMM and the DNN-SGMM systems with very small and large number of hidden layers have resulted in slightly degraded speech recognition performances, when compared with the moderate number of hidden layers. This degradation is attributed to the elementary and the complex non-linear mapping between the input and the output of DNN. The large number of hidden layers increase the parameters for estimation for the DNN-HMM and the DNN-SGMM systems, which results in depleted speech recognition performances for the stressed speech. Whereas, the increased WERs for less number of hidden layers acknowledge about the inefficacy of the DNN to model the posterior probabilities over the states of GMM-HMM and SGMM-HMM systems. The incorporation of discriminative capability of DNN with adequate number of hidden layers into the dynamic time warping property of GMM-HMM and SGMM-HMM systems has resulted in a fascinating modelling technique for capturing the acoustic variations in the stressed speech resulting from stress.

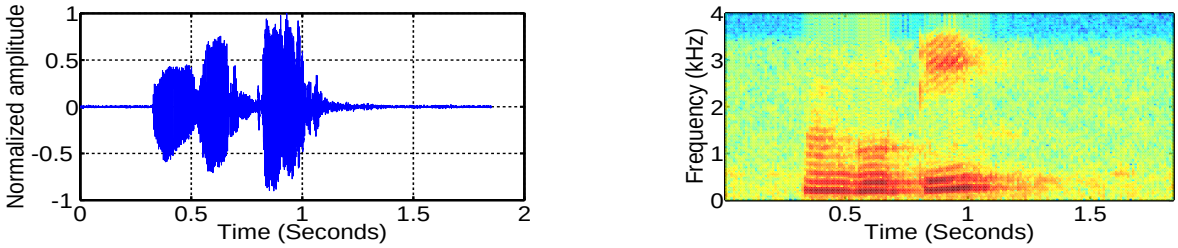
All the acoustic modelling techniques studied in this work are found to result in great speech recognition performances with very low WER values, when models are trained using the features of neutral speech data and tested also using the features of neutral speech utterances. Whereas, the significantly increased values of WER are reported in Table 2.1-Table 2.3 for the recognition of stressed speech utterances. These experimental results demonstrate the similar phonetic structure for the training and the testing environments of speech recognition system, when trained and tested using the neutral speech. The alteration in the speech production system under stress condition creates a very high variance between the phonetic structure of neutral and stressed speech. Con-

sequently, the ASR system trained on the neutral speech, when tested using the stressed speech exhibits severe degradation in the recognition performances. The acoustic mismatch between the neutral and the stressed speech generates a dissimilar training and test environments for the speech recognizer. This degrades the accuracy of speech recognition for recognizing the stressed speech.

2.4 Motivation for the Present Investigation

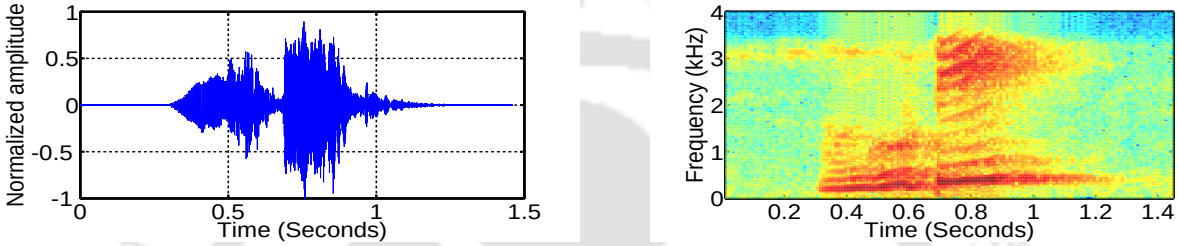
In our everyday experience, stress has salient impact on the speech production system. Speech produced under stress condition has different characteristics compared to the speech produced under neutral or normal condition. The properties of acoustic parameters of speech signals such as formants, pitch, speaking rate, intensity, energy, glottal pulses etc., vary significantly across the speech produced under neutral and stress conditions. These acoustic mismatches are often found to be the main causes for the degradation in the performances of most of the real life, the large scale laboratory and the commercial applications, which involve their interaction with machines for the users under stress conditions [8, 9, 11–18]. The normalization of stress-specific attributes helps in accomplishing the user-affable practical life applications. The stress normalization is progressively appealing concern from the wider range of researchers involved in the area of stressed speech.

To observe the acoustic dissimilarities, differences between the characteristics of neutral and stressed speech have been investigated in time and frequency domains. The time domain investigation measures the variation in the air pressure that human auditory systems are able to perceive as sound with respect to the time. In frequency domain, the interpretation of speech signal has been accomplished by studying the changes in the spectral distribution over the time parameter. Variations in the air pressure and the spectral contents of speech signals are quantified by analyzing their waveforms and spectrograms, respectively. The speech utterances of two different words /angoothi/ and /daakghar/ of SUSSC database [118] recorded from the same speaker under neutral and four stress conditions namely: angry, sad, lombard and happy, respectively, have been employed for measuring the acoustic discrepancies. Figure 2.8 and Figure 2.9 show the waveforms of speech signals and their spectrograms of both the words. Visible differences can be observed in the duration and the spectral energy of speech utterances from their waveforms and spectrograms, respectively. The speech uttered with sad and lombard conditions have durations more than that uttered with neutral condition as depicted in Figure 2.8(a), Figure 2.8(e), Figure 2.8(g), Figure 2.9(a), Figure 2.9(e) and



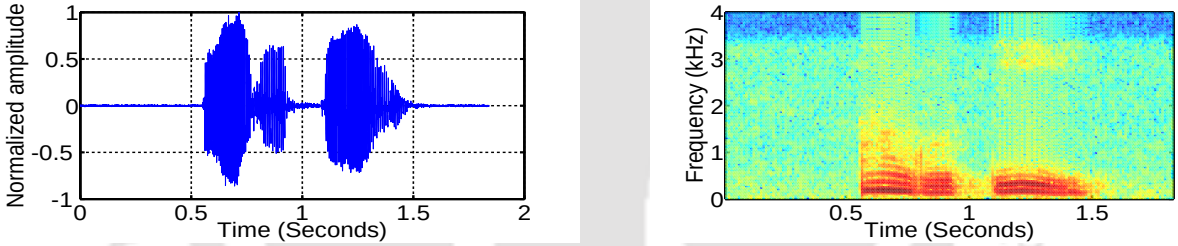
(a) (b)

Speech recorded under neutral condition: (a) waveform and (b) spectrogram



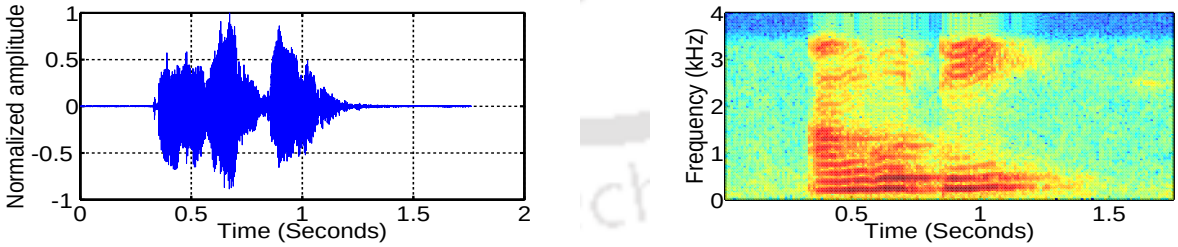
(c) (d)

Speech recorded under angry condition: (c) waveform and (d) spectrogram



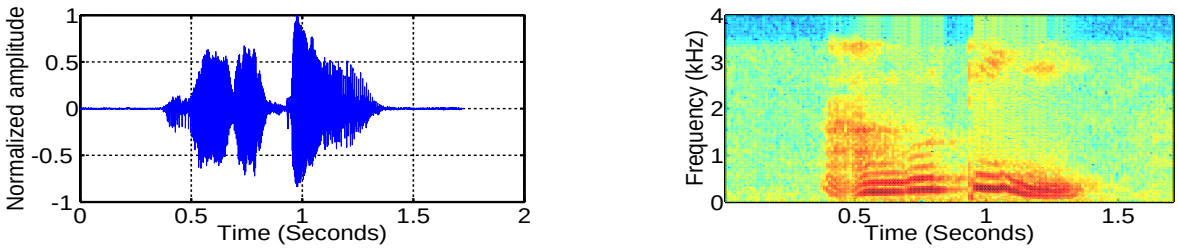
(e) (f)

Speech recorded under sad condition: (e) waveform and (f) spectrogram



(g) (h)

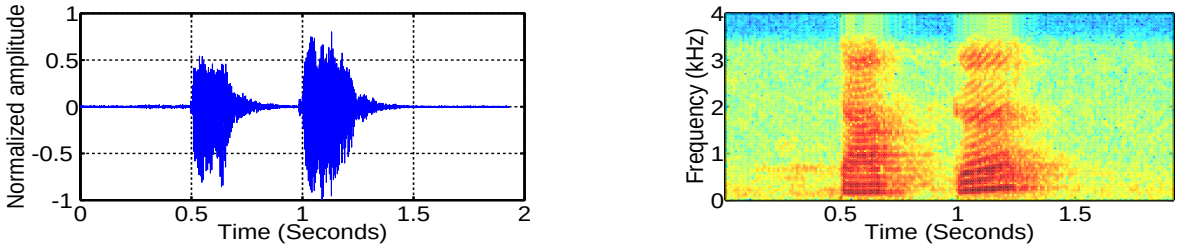
Speech recorded under lombard condition: (g) waveform and (h) spectrogram



(i) (j)

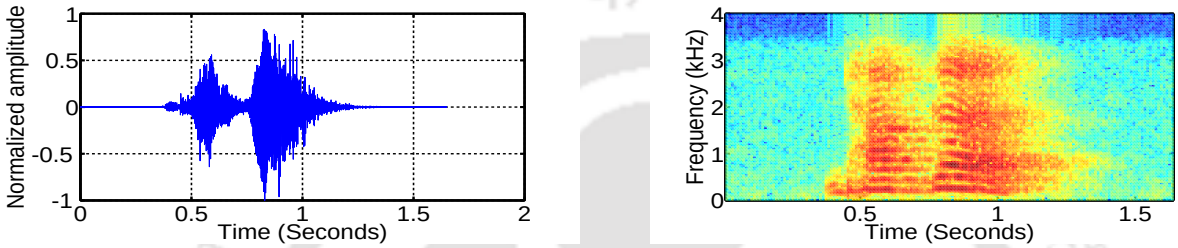
Speech recorded under happy condition: (i) waveform and (j) spectrogram

Figure 2.8: The visual analysis of speech utterances of word /angoothi/ recorded under neutral, angry, sad, lombard and happy conditions by plotting their waveforms and spectrograms, respectively.



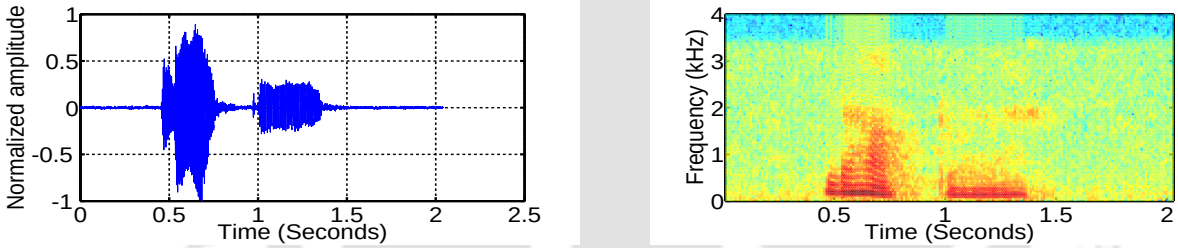
(a) (b)

Speech recorded under neutral condition: (a) waveform and (b) spectrogram



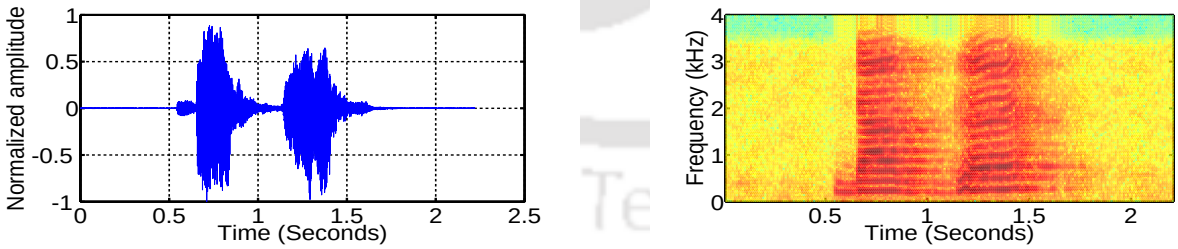
(c) (d)

Speech recorded under angry condition: (c) waveform and (d) spectrogram



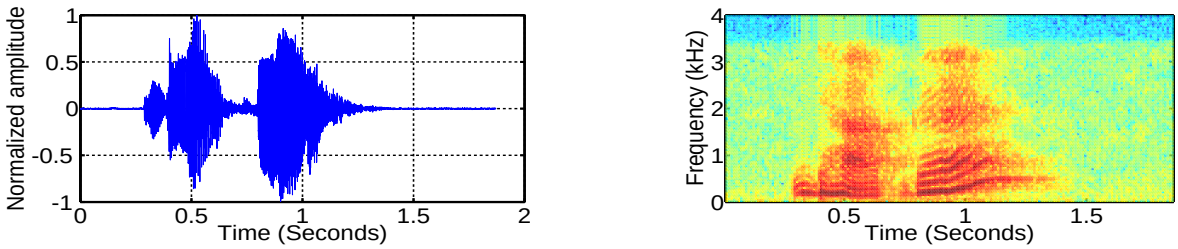
(e) (f)

Speech recorded under sad condition: (e) waveform and (f) spectrogram



(g) (h)

Speech recorded under lombard condition: (g) waveform and (h) spectrogram



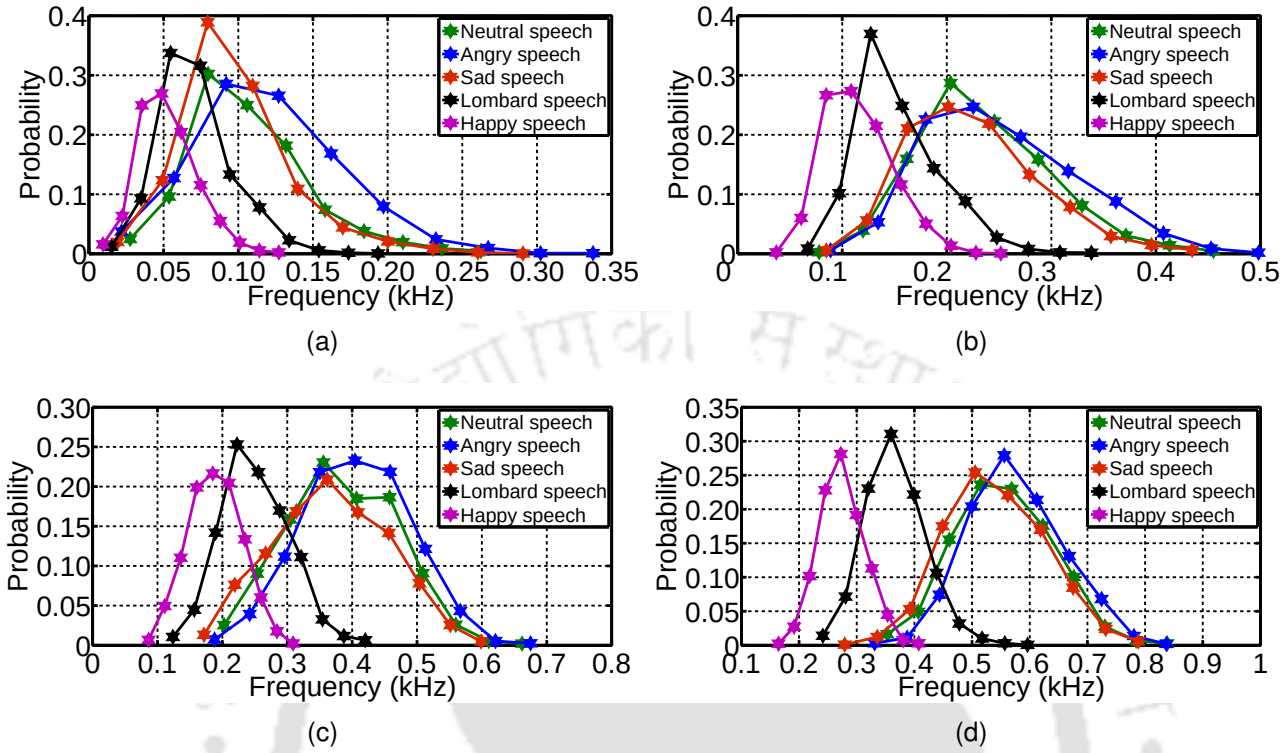
(i) (j)

Speech recorded under happy condition: (i) waveform and (j) spectrogram

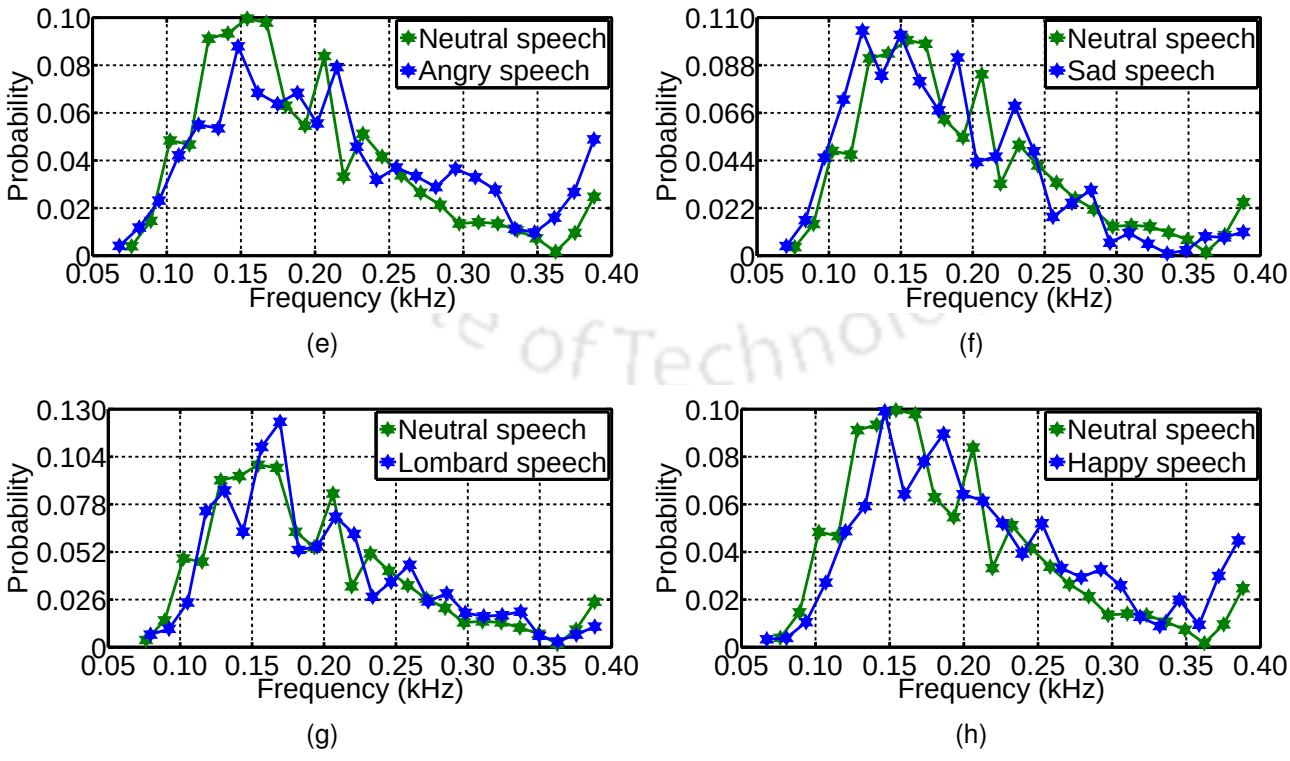
Figure 2.9: The visual analysis of speech utterances of word /daakghar/ recorded under neutral, angry, sad, lombard and happy conditions by plotting their waveforms and spectrograms, respectively.

Figure 2.9(g). The total energy of speech signals has a higher value in cases of speech produced under angry, lombard and happy conditions compared to neutral and sad conditions. The spectrograms illustrate that, the relative energy in lower frequencies compared to higher frequencies has higher value in case of neutral speech than the speech uttered with sad emotion as depicted in Figure 2.8(b), Figure 2.8(f), Figure 2.9(b) and Figure 2.9(f). The spectral energy has higher values in the speech signal uttered in angry and lombard conditions, when compared to the speech produced under neutral, sad and happy conditions. These observations demonstrate that, the characteristics of the speech signal changes under different stress conditions. The duration of speech signal vary significantly under stress condition, when compared to the neutral condition. The spectral energy associated with each frequency band is different under different stress conditions. The changes in the air pressure and the spectral distribution with respect to the time as available in the literature, have considered very effective in analyzing the stressed speech. These studies motivate us to explore the duration parameter and the spectral energy for diminishing the divergences resulting from stress.

Stress induces variabilities in the speech production system and it results in changed characteristics for the vocal-tract system and the excitation source. In literature, several research works have been reported on the changes in the formant and the pitch frequencies [4–6]. In this work, the inconsistencies in the formant and the pitch frequencies under neutral and stress conditions are studied by conducting an experimental study using the SUSSC database reported in [118]. The frequency values (F_1, F_2, F_3 and F_4) corresponding to the first four formants and the pitch frequency F_0 are analyzed by estimating their probability densities. To measure the acoustic mismatch, approximately 2000 frames of speech utterances of two distinct words namely: /angoothi/ and /daakghar/, respectively, recorded by different speakers have been employed. Figure 2.10 and Figure 2.11 have shown the probability density plots for the formant and the pitch frequencies for the utterances of both the explored words. It is evident that, the vocal-tract system has non-linearly (closer to Gaussian characteristics) increasing formants of shifted and scaled attributes under stress conditions compared to the neutral condition. These characteristics are exhibited for all the studied words and stress classes as depicted in Figure 2.10(a)$-$Figure 2.10(d) and Figure 2.11(a)$-$Figure 2.11(d). The observations of this experiment also revealed that, the formant frequencies for the lombard and the happy speech are located in the low frequency region than the formant frequencies for the neutral, the angry and the sad speech. These shifted and scaled frequency values of formants are mainly attributed to the

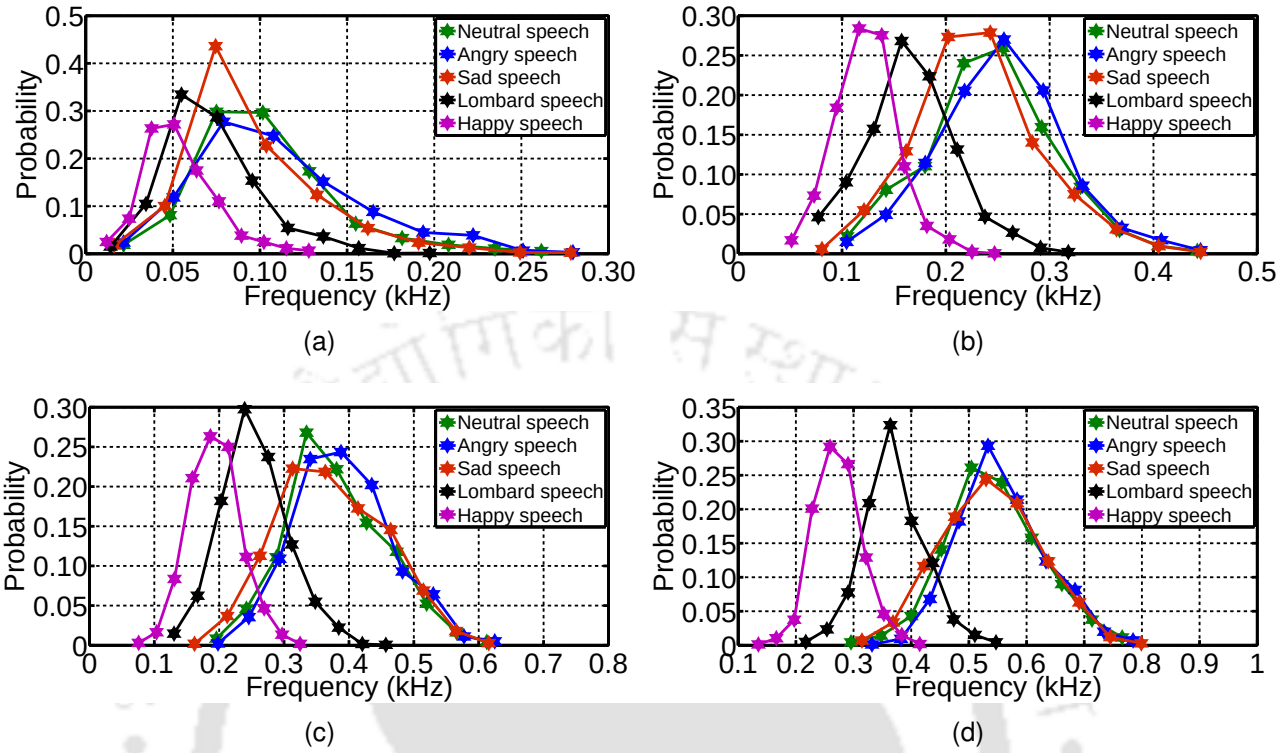


Formant frequencies; (a) first formant frequency, F_1 (b) second formant frequency, F_2 (c) third formant frequency, F_3 and (d) fourth formant frequency, F_4

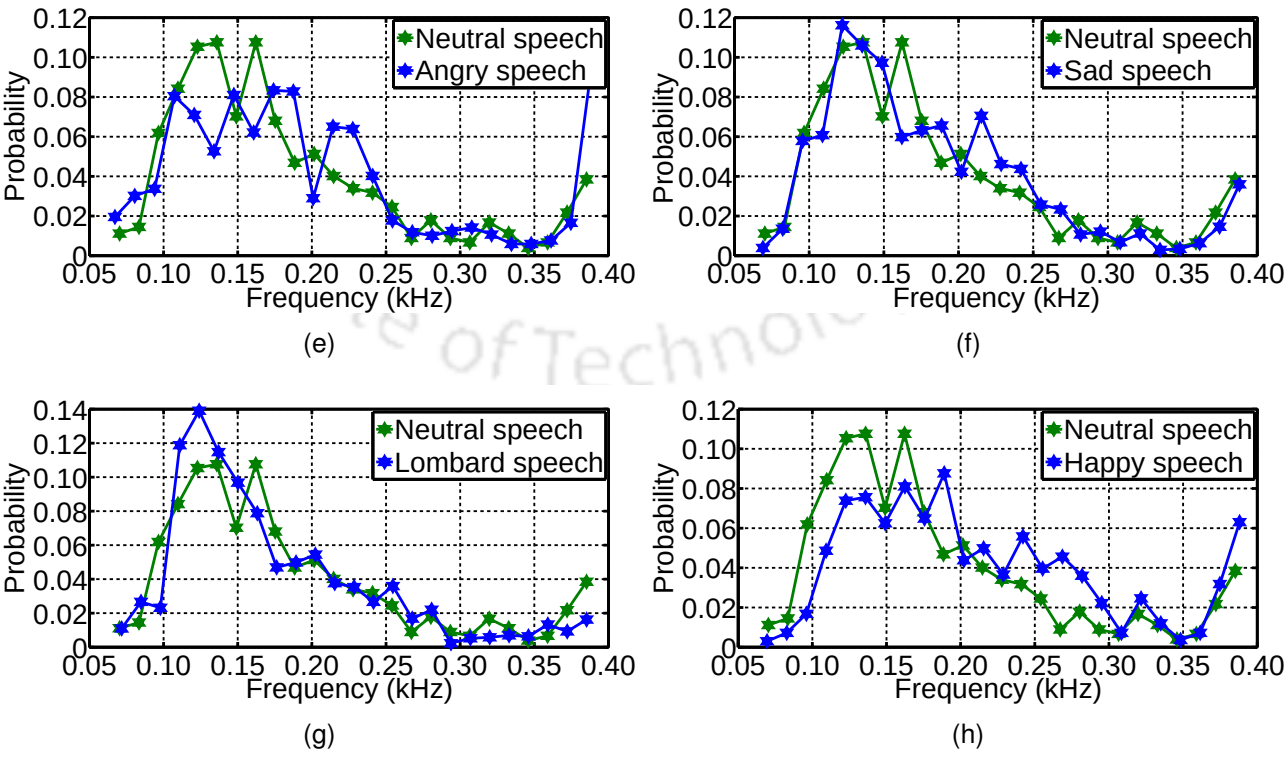


Pitch frequency (F_0); (e) angry, (f) sad, (g) lombard and (h) happy

Figure 2.10: The probability density plots for the formant and the pitch frequencies over the utterances of word 'angootha' produced under neutral and four stress cases studied in this work.



Formant frequencies; (a) first formant frequency, F_1 (b) second formant frequency, F_2 (c) third formant frequency, F_3 and (d) fourth formant frequency, F_4



Pitch frequency (F_0); (e) angry, (f) sad, (g) lombard and (h) happy

Figure 2.11: The probability density plots for the formant and the pitch frequencies over the utterances of word /daakghar/ produced under neutral and four stress cases studied in this work.

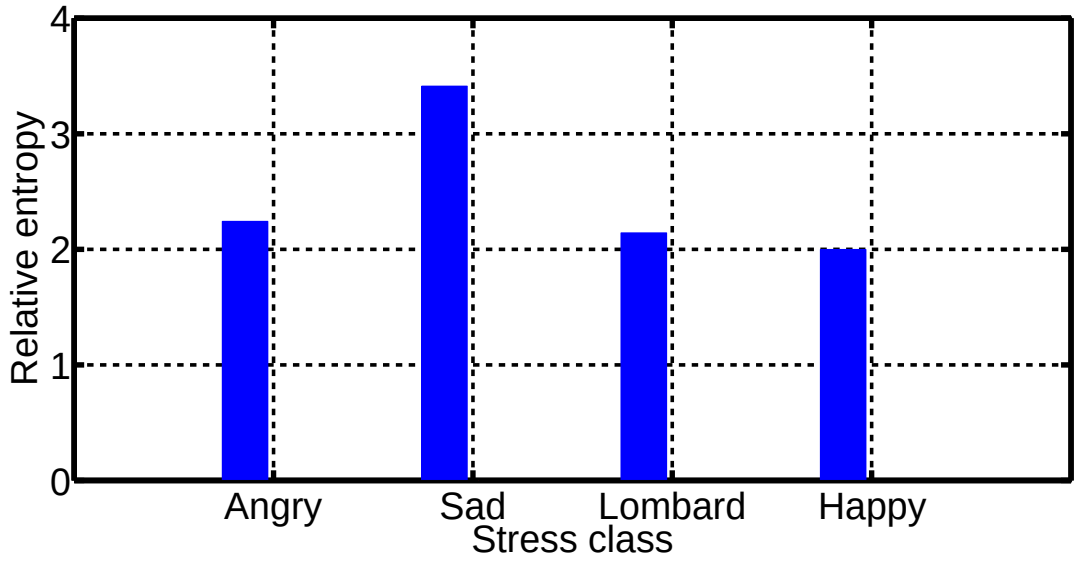


Figure 2.12: The relative entropy between the Gaussian-subspaces developed using the neutral and stressed speech features by exploring the Kullback Leibler (KL) divergence metric.

movement of vocal-tract articulators to configure the vocal-tract shape resulting in a non-linear airflow pattern for perception of the additional information about stress. The pitch frequency (F_0) for the speech utterances under all the explored stress classes and words followed the different path than the neutral speech as shown in Figure 2.10(e)−Figure 2.10(h) and Figure 2.11(e)−Figure 2.11(h). The divergences in probability distributions for the pitch frequency under neutral and stress conditions give information about the anatomical and the physiological differences among the speakers. Speakers with age and gender variation have different periodicity of the excitation source for the production of stressed speech. These experimental results acknowledge the changes in the characteristics of the vocal-tract system and the excitation source under stress condition. The modification in the vocal-tract system and the excitation source under stress condition creates a high overlap between the different speech units of stressed speech and it introduces a large acoustic mismatch between the patterns of neutral and stressed speech. Therefore, the investigation on the modification in the vocal-tract system under stress condition can help in reducing the acoustic mismatch between the neutral and the stressed speech signals.

In this study, the variance mismatch between the neutral and the stressed speech is further investigated by evaluating the similarity between the Gaussian-subspaces developed on the SUSSC database reported in [118]. In order to quantify the acoustic mismatch, the Kullback Leibler (KL) diver-

gence metric has been explored to measure the relative entropy between the Gaussian-subspaces [140, 141]. The relative entropy, $D_{\mathrm{KL}}(f(x), g(x)) = \int f(x) \log \frac{f(x)}{g(x)} d(x)$ measures the similarity between two probability density functions $f(x)$ and $g(x)$ and satisfies the following three properties: (i) $D_{\mathrm{KL}}(f(x), f(x)) = 0$, referred to as the self similarity, (ii) $D_{\mathrm{KL}}(f(x), g(x)) = 0$, only if $f(x) = g(x)$ called as the self identification, (iii) $D_{\mathrm{KL}}(f(x), g(x)) \geq 0$, only if $f(x) \neq g(x)$, respectively. To implement this experiment, 495 speech utterances are collected by 15 speakers (5 female and 10 male) under neutral and each four stress classes namely: angry, sad, lombard and happy, respectively. For computing speech features, the speech utterances are analyzed using a 20 msec Hamming window and frame shift of 10 msec. A 21-channel Mel-filterbank is employed to determine the 13-dimensional base MFCC $(C0 - C12)$ features [100]. A separate GMM, constituting the mixture of 64 Gaussian densities is trained using the MFCC features of the neutral speech and the speech produced under four stress conditions explored in this work. The subspaces corresponding to these GMMs are referred to as the Gaussian-subspaces, which constitute the information about the different stress conditions. The relative entropy between the Gaussian-subspaces developed using the neutral and the stressed speech are summarized using the bar plot as shown in Figure 2.12. Significant values of relative entropy with KL divergences of 2.242, 3.412, 2.142 and 1.999 are obtained between the GMMs trained on the neutral speech and the speech produced with angry sad, lombard and happy conditions, respectively. The reported KL divergence values revealed that, the Gaussian-subspace corresponding to the sad speech exhibits maximum distance with relative entropy value of 3.412 among all other studied stress classes from the Gaussian-subspace created using the neutral speech. Whereas, the Gaussian-subspace created using the happy speech has less dissimilarities with a relative entropy of 1.999 compared to all other explored stress classes from the Gaussian-subspace developed using the neutral speech utterances. These reported KL divergences illustrate that, the Gaussian-subspace developed using the stressed speech exhibits the dissimilar characteristics compared to the Gaussian-subspace created using the neutral speech. These dissimilarities between the Gaussian-subspaces manifest that, stress alters the characteristics of speech signals and it increases the variance mismatch between the neutral and the stressed speech.

These experimental observations presented demonstrate that, stress causes a large scale of acoustic mismatch between the different speech units of neutral and stressed speech signals. During the last few decades, though some research publications in the area of stressed speech have ap-

peared, there is little effort for the normalization of stress information. In most cases, changes in the acoustic properties of stressed speech have been investigated on an assumption that, the stressed speech has linear characteristics. On the other hand, few studies have exploited the non-linear properties of stressed speech. Stress induces non-linearity in the pattern of stressed speech compared to the pattern of neutral speech. It would be interesting to investigate the non-linear characteristics of stressed speech by exploring the subspace projection-based approaches. Normalization of stress information by investigating the non-linear properties may help in establishing the fact whether speech and stress information are related linearly or non-linearly. In literature, the investigation on the changes in the vocal-tract system and the excitation source are reported to be very effective for the robust representation of stressed speech units [25, 43–45]. The experimental studies summarized in Figure 2.10– Figure 2.11 show the changes in the frequency values for the formant and the pitch under stress condition. The formant- and pitch-dependent distortions have resulted in high variance values for the spectral features as depicted in Figure 2.3– Figure 2.4. These observations motivate us for the investigation on the modification in the vocal-tract system under stress condition and it can lead to the development of effective stress normalization techniques. Conventionally, the stressed speech is analyzed by the studying the changes in the characteristics in the feature- and model-space. Since, the anatomical and the physiological characteristics of speech production system carry the different properties for different speakers. A large acoustic variation is noted between the speech signals produced by different speakers under similar stress conditions [24,33,47,48,58,59,61,81,82,87–92,98]. In these studies, the analysis on the changes in the speech production system due to speakers were found to be effective for normalizing the speaker variability. The suppression of speaker variability can help in improving the robustness of stress normalization techniques.

With the development of various signal processing and pattern matching techniques, in most of the research interests including stressed speech, the subspace projection is adopted as one of the state-of-the-art techniques for the separation of the desired signal from the undesired signal [142]. The literature reviews summarized in Section 1.3 in Chapter 1 have shown the effectiveness of subspace projection-based approaches for the analysis of stressed speech. The efficiency of subspace projection depends on the following two factors.

(i) The learning mechanism of subspace projection matrix.

(ii) The technique employed for the subspace projection.

Speech information is present in both the neutral and the stressed speech signals. Therefore, the learning of subspace projection matrix capturing the principal dimensions of the acoustic variations represented by the neutral speech can be experimented for reducing the consequence of stress as well as for retaining the speech-specific attributes. From the reported works in [19, 41], it is observed that, for the stressed speech, the high frequency region is affected more compared to the low frequency region [19, 41]. Therefore, lowering the rank of the projection matrix may, in turn, reduce the mismatch in the variances resulting from the stress and motivate us for low-rank subspace projection for further analysis of stressed speech in lower dimensional subspace.

Motivated by these facts, this thesis documents our investigations on the normalization of stress information by exploring the novel speech subspace modelling with speaker adaptation techniques. The work presented in this dissertation is broadly divided into following five major contributions.

(i) The linear and the non-linear subspace modelling approaches have been explored to investigate the characteristics of stressed speech. The stress information is normalized by the projection of stressed speech onto a subspace, which consists of properties of neutral speech. In this work, the linear subspace has been developed using the orthogonal projection and linear transformation techniques. The non-linearity between the speech- and the stress-specific attributes are investigated by exploring the subspace projection onto the non-linear data space. An effort is made to reduce the acoustic mismatch between neutral and stressed speech by accomplishing the subspace projection through the non-linear transformation using the polynomial function. This study investigates the non-linearity between the speech and the stress information by varying the order of polynomial function.

(ii) The vocal-tract system is studied in the Gaussian-subspace for the normalization of stress-specific attributes. The parameters of vocal-tract system for the neutral and the stressed speech are projected onto a common Gaussian-subspace, which constitutes the vocal-tract system parameters of neutral speech. In this work, the subspace projection is derived using the posterior probability information and it has resulted in the posteriorgram features. Both the neutral and the stressed speech are synthesized using the corresponding projected vocal-tract system parameters and considered as speech signals comprising similar acoustic properties.

(iii) The alteration in the characteristics of vocal-tract system under stress condition is further investigated in the sparse domain. A novel sparse representation of vocal-tract system param-

eters of neutral and stressed speech is explored over the dictionary comprising the vocal-tract system parameters of neutral speech data. The experimental studies summarized in in Figure 2.8−Figure 2.9 illustrate the alteration in the duration of speech utterances produced under stress conditions compared to the neutral condition. Speakers change the speaking rate to emphasize or deemphasize the information about the stress environment. These observations motivate us to introduce the information about the duration parameter of speech utterances for the development of utterance-specific adaptive dictionary. The information about the duration parameter is modeled using the K-nearest-neighbour (K-NN) algorithm-based non-parametric probability density estimation method [143]. Also, the learning of invariable size global dictionary by employing the K-SVD algorithm [76] is carried out to compare the relative effect of stress normalization using the utterance-specific adaptive dictionary.

(iv) The thesis work is also dedicated to search for a subspace that consists of speech information along with the suppressed speaker variability. The feature space maximum-likelihood linear regression (fMLLR) method is exploited in the speaker adaptive training (SAT) framework to normalize the speaker information [95, 97, 144–146]. Furthermore, the effectiveness of all the proposed stress normalization methods are investigated in the lower dimensional subspace. The linear discriminant analysis (LDA) [147] and the heteroscedastic linear discriminant analysis (HLDA) [148–150] have been explored in the maximum likelihood linear transformation (MLLT)-based semi-tied adaptation technique for the low-rank subspace projection.

(v) In modelling paradigm, the automatic speech recognition (ASR) systems developed by using the Gaussian mixture model (GMM) [1], the subspace Gaussian mixture model (SGMM) [52–54] and the deep neutral network (DNN) [55, 56] acoustic modelling techniques have been explored to measure the effectiveness of proposed stress normalization methods. In addition to these, the visual analysis and the error analysis are also presented to validate the effectiveness of proposed stress normalization techniques. The visual analysis measures the variations in the air pressure and the spectral properties over the time by studying the waveform and the spectrogram, respectively. In the error analysis, we have exploited the relative entropy between the Gaussian-subspaces by introducing the Kullback Leibler (KL) divergence metric.

3

Linear and Non-Linear Subspace Modeling for Stress Normalization

Contents

This chapter explores the linear and the non-linear characteristics of speech produced under stress condition. As discussed in Chapter 1 and Chapter 2, the acoustic mismatch between the neutral and the stressed speech affects many practical applications in real life [8, 9, 11–18]. The development of human-computer interaction (HCI) system is one of the necessities for implementing theses practical applications. The affective computing of HCI system depends on the robust automatic speech recognition (ASR) system [23, 24, 58, 59]. For the development of robust ASR system, the training and the test environments should comprise of similar acoustic properties. The ASR system trained using the neutral speech data exhibits the degraded performances against the users under stress conditions. The acoustic variations between the neutral and the stressed speech create mismatched training and testing conditions for the ASR system. The normalization of stress-specific attributes can help in reducing the acoustic mismatch between the speech units of neutral and stressed speech. In literature, numerous approaches have been reported for normalizing the stress information [6, 19, 20, 27, 33, 64]. The subspace projection is one of the techniques employed for normalizing the stress-specific attributes [27, 40, 45–49, 51, 69, 72]. The projection of stressed speech onto a subspace consisting of speech produced under neutral condition, which is called as the speech subspace can result in normalizing the stress information. The efficiency of stress normalization will depend on the learning mechanism of effective speech subspace and the technique employed for the accomplishment of subspace projection. In literature, various methods, exploiting the linear and the non-linear characteristics have been developed for the subspace modelling [142].

In this work, the linear and the non-linear properties of stressed speech have been investigated in the linear and the non-linear subspace for designing the effective stress normalization algorithms, respectively. The proposed approach has incorporated the orthogonal projection and the non-linear transformation techniques to develop the linear and the non-linear subspace, respectively. To further improve the performances of stress normalization methods, the proposed subspace modelling techniques have been exploited into the another subspace consisting of decorrelated properties. Moreover, the dimension of this subspace is varied by introducing the low-rank subspace projection method for studying the consequence of stress in different frequency bands. The heteroscedastic linear discriminant analysis (HLDA) is employed in maximum likelihood linear transformation (MLLT)-based semi-tied adaptation technique to adapt the feature- and the model-space onto the decorrelated subspace [148–150]. In this chapter, we have presented the analysis and the evaluation of

proposed stress normalization techniques using the TEO-CB-Auto-Env features [44, 45] and the Mel-frequency cepstral coefficients (MFCCs) [100] features under four stress conditions namely: angry, sad. lombard and happy, respectively.

The remainder of this Chapter is organized as follows: The proposed orthogonal projection-based linear subspace modelling approach for normalizing the stress-specific attributes is presented in Section 3.1. In Section 3.2, the methodology of low-rank subspace projection is described. Section 3.3 contains the descriptions on the experimental setup used followed by the results and the discussion of the proposed linear subspace modelling technique. The transformation-based non-linear subspace modelling technique for the development of effective stress normalization algorithm is proposed in Section 3.4. The experimental evaluations for the non-linear subspace modelling approach are presented in Section 3.5. Finally, the findings of these works are summarized in Section 3.6.

3.1 Linear Subspace Modeling Using Orthogonal Projection

Stressed speech comprises the stress-specific attributes and the phonetic information also referred to as the speech-specific attributes. Consequently, the stressed speech differs considerably from the neutral speech in many aspects and characteristics, thus resulting in a high acoustic variability. It is hypothesized that, the speech- and the stress-specific attributes of stressed speech are linearly related to each other. To exploit this linear relationship, a novel linear subspace modelling approach using the orthogonal projection-based subspace filtering technique has been explored for the normalization of stress-specific divergences. Stress information is suppressed by projecting the stressed speech orthogonally onto a subspace, which consists of acoustic properties of speech produced under neutral condition, generally called as the speech subspace. The creation of proposed speech subspace using the neutral speech features can help in reducing the acoustic mismatch between the neutral and the stressed speech. The effectiveness of the proposed method depends on the following two factors namely: the filtering of an effective speech subspace and the mechanism of orthogonal projection of stressed speech onto the filtered effective speech subspace, respectively. The orthogonal projection of stressed speech features onto the speech subspace can help in normalizing the stress-specific attributes. The orthogonally projected stressed speech features are considered as the stress normalized features and create the linear subspace, whose bases effectively span the speech-specific attributes. The proposed stress normalization approach is described in the block

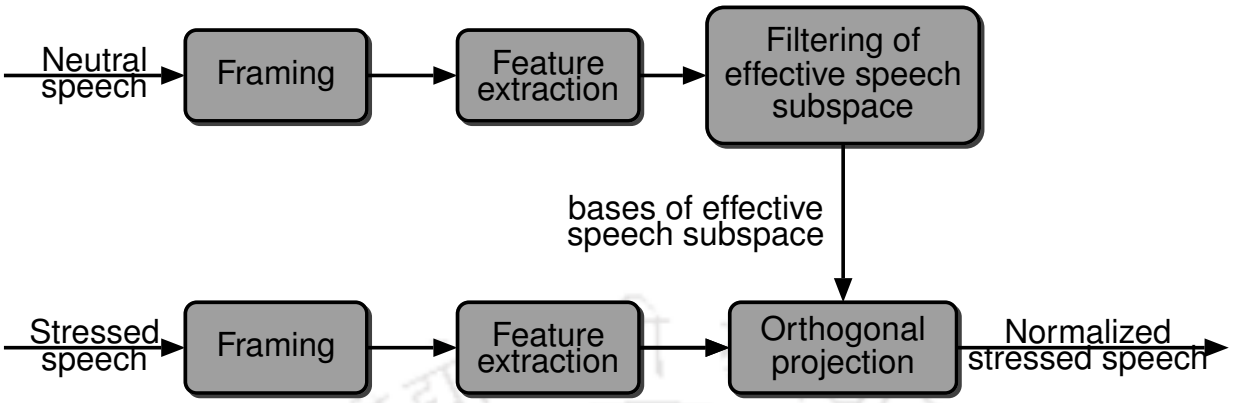


Figure 3.1: The proposed linear subspace modelling technique for stress normalization.

diagram shown in Figure 3.1. As depicted in this block diagram, the stressed speech features are orthogonally projected onto the bases of effective speech subspace for normalizing the stress information. In the following, we have described the methodology of estimation of basis vectors of speech subspace. A detail discussion on the filtering of an effective speech subspace is also presented. This is followed by the normalization of stress information using the filtered effective speech subspace.

3.1.1 Proposed Speech Subspace

The subspace, which consists of acoustically rich phonetic information or speech-specific attributes is called as speech subspace. As discussed earlier, speech information is present in both the neutral and the stressed speech signals. Therefore, the learning of speech subspace using the neutral speech data can help in estimating the bases, which span the speech-specific attributes. The orthogonal projection of stressed speech features onto the speech subspace decomposes it into two components namely: the orthogonal projection component and the complement orthogonal projection component, respectively. The orthogonal projection component of the stressed speech features belongs to the column space of the speech subspace. Consequently, it comprises of speech-specific attributes and reduces variances resulting from stress. In this work, the speech subspace is learned using the neutral speech utterances by exploiting the K-means clustering [1] followed by the singular value decomposition (SVD) [142] techniques as depicted in block diagram shown in Figure 3.2. The following steps summarize the method for the estimation of bases of the proposed speech subspace.

Step I: At first, feature vectors are extracted using all the neutral speech utterances of train-

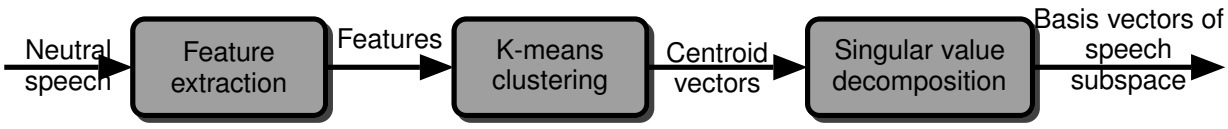


Figure 3.2: Estimation of basis vectors of the proposed speech subspace.

ing database, which belong to word w. The training database comprises the neutral speech utterances having total W, $(1 \leq w \leq W)$ words.

Step II: In the next step, a set of representative vectors (mean vectors or centroids), called as the codebook are determined for all the feature vectors corresponding to word w using the K-means clustering technique. The set of representative vectors $\{\mathbf{c}_r^w\}_{r=1}^R$ are arranged in codebook matrix $\mathbf{C}^w = \begin{bmatrix} \mathbf{c}_1^w & \mathbf{c}_2^w & \cdots & \mathbf{c}_R^w \end{bmatrix}$. The size of codebook matrix $\mathbf{C}^w$ is $K \times R$. Where K and R denote the dimension of feature vector and the number of centroid vectors, respectively.

Step III: The independency property of basis vectors is obtained by performing the SVD on the codebook matrix $\mathbf{C}^w$ and determined as follows,

$$\mathbf{C}^w = \mathbf{U}^w \mathbf{S}^w \mathbf{V}^{wT} \tag{3.1}$$

where $\mathbf{U}^w$ and $\mathbf{V}^w$ are $K \times K$ and $R \times R$ orthonormal matrices, respectively. $\mathbf{S}^w$ is $K \times R$ diagonal matrix. The operator "T" represents the transpose operation. The set of columns $\{\mathbf{u}_j^w\}_{j=1}^K$ of the matrix $\mathbf{U}^w$ are a set of independent vectors since they are mutually orthonormal and can span the K-dimensional subspace [142].

Step IV: In this work, different sets of orthonormal column vectors $\{\mathbf{u}_j^w\}_{j=1}^K$ are tested as the basis vectors of speech subspace. The eigenvalues corresponding to each column vectors $\{\mathbf{u}_j^w\}_{j=1}^K$ are exploited for determining the basis vectors as follows: At first, it is assumed that one column vector $\mathbf{u}_j^w$ corresponding to the maximum eigenvalue can span the speech subspace, which consists of speech-specific attributes of stressed speech. After that two column vectors corresponding to the two highest eigenvalues are tested as the basis vectors for the speech subspace. In this way, all the column vectors $\{\mathbf{u}_j^w\}_{j=1}^K$ are used as the basis vectors of speech

subspace with decreasing eigenvalues given as,

$$\mathbf{U}_1^w = \begin{bmatrix} \mathbf{u}_1^w \end{bmatrix}$$
$$\mathbf{U}_2^w = \begin{bmatrix} \mathbf{u}_1^w & \mathbf{u}_2^w \end{bmatrix}$$
$$\vdots$$
$$\mathbf{U}_K^w = \begin{bmatrix} \mathbf{u}_1^w & \mathbf{u}_2^w & \cdots & \mathbf{u}_K^w \end{bmatrix}$$

The eigenvalues corresponding to the orthonormal column vectors $\mathbf{u}_1^w, \mathbf{u}_2^w, \cdots, \mathbf{u}_K^w$ are denoted by $\sigma_1^w, \sigma_2^w, \cdots, \sigma_K^w$, respectively. These eigenvalues are arranged in decreasing order of their magnitude i.e., $\sigma_1^w > \sigma_2^w > \cdots > \sigma_K^w$. The orthonormal vectors in $\mathbf{U}_1^w, \mathbf{U}_2^w, \cdots, \mathbf{U}_K^w$ are tested as the basis vectors of speech subspaces, which are the different sets of orthonormal column vectors of $\mathbf{U}^w$. In this manner, the sets of basis vectors $\{\mathbf{U}_j^w\}_{j=1}^K$ are determined for each word w of neutral speech utterances present in the training data.

3.1.2 Filtering of an Effective Speech Subspace

As discussed in previous Subsection 3.1.1, the sets of basis vectors $\{\mathbf{U}_j^w\}_{j=1}^K$ are determined for each word w of the neutral speech data (training data). Therefore, for any given fixed number of basis vectors j between $1 \leq j \leq K$, the filtering of an effective speech subspace for the test stressed speech utterance depends on the particular word w. The filtered effective speech subspace spans the speech-specific attributes and helps in normalizing the stress information of the stressed speech utterance. The following steps describe the method for filtering of an effective speech subspace.

Step I: The given test stressed speech utterance is parametrized into the set of feature vectors $\{\mathbf{v}_f\}_{f=1}^F$. Where, F represents the total frames of stressed speech utterance. The orthogonal projection component $\mathbf{v}_f{}_p^w$ of the stressed speech vector $\mathbf{v}_f$ onto the speech subspace U_j^w, which corresponds to set of basis vectors $\mathbf{U}_j^w = \begin{bmatrix} \mathbf{u}_1^w & \mathbf{u}_2^w & \cdots & \mathbf{u}_j^w \end{bmatrix}$, is determined as follows

$$\mathbf{v}_f{}_p^w = (\mathbf{v}_f \cdot \mathbf{u}_1^w)\mathbf{u}_1^w + (\mathbf{v}_f \cdot \mathbf{u}_2^w)\mathbf{u}_2^w + \cdots + (\mathbf{v}_f \cdot \mathbf{u}_j^w)\mathbf{u}_j^w \tag{3.2}$$

where, the operator "." represents the dot product between the vectors. The complement or-

thogonal projection component $\mathbf{v}_{f_c}^{w}$ is determined as

$$\mathbf{v}_{f_c}^{w} = \mathbf{v}_f - \mathbf{v}_{f_p}^{w} \tag{3.3}$$

Step II: In the next step, the average value of L_2 norm $\|.\|$ of complement orthogonal projection components $\{\mathbf{v}_{f\ c}^{w}\}_{f=1}^{F}$ over all the frames is calculated as,

$$E_c^w = \frac{1}{F}\left[\|\mathbf{v}_{1_c}^{w}\| + \|\mathbf{v}_{2_c}^{w}\| + \cdots + \|\mathbf{v}_{F_c}^{w}\|\right] \tag{3.4}$$

Step III: Following step I and step II, the value of E_c^w for stressed speech utterance is determined using all the speech subspaces $\{U_j^w\}_{w=1}^{W}$ corresponding to all the words of neutral speech data. The stressed speech utterance belongs to the speech subspace $U_j^{w'}$ for which the value of E_c^w would be minimum given as

$$w' = \underset{w}{\operatorname{argmin}}(E_c^w) \tag{3.5}$$

The subspace $U_j^{w'}$, which corresponds to the basis vectors $\mathbf{U}_j^{w'}$ is considered as the filtered effective speech subspace for the given test stressed speech utterance.

3.1.3 Stress Normalization Using the Filtered Effective Speech Subspace

The stress-specific divergences are normalized by projecting the feature vectors $\{\mathbf{v}_f\}_{f=1}^{F}$ of stressed speech utterance orthogonally onto the filtered effective speech subspace $U_j^{w'}$. The set of orthogonal projection components $\{\mathbf{v}_{f_p}^{w'}\}_{f=1}^{F}$ are considered as the normalized stressed speech features and are determined using the Eq. 3.2 as follows

$$\mathbf{v}_{f_p}^{w'} = (\mathbf{v}_f \cdot \mathbf{u}_1^{w'})\mathbf{u}_1^{w'} + (\mathbf{v}_f \cdot \mathbf{u}_2^{w'})\mathbf{u}_2^{w'} + \cdots + (\mathbf{v}_f \cdot \mathbf{u}_j^{w'})\mathbf{u}_j^{w'} \tag{3.6}$$

These normalized stressed speech features as the orthogonal projection components $\{\mathbf{v}_{f_p}^{w'}\}_{f=1}^{F}$ have exploited the linearity between the speech- and stress-specific attributes and develop the linear-subspace. Consequently, the normalized stressed speech features belong to the column space of this linear subspace, whose bases span the speech information of stressed speech. Figure 3.3 demonstrates the stress normalization by exploring the proposed orthogonal projection of stressed speech feature vector $\mathbf{v}_f$ onto the set of basis vectors $\mathbf{U}_j^{w'} = \begin{bmatrix}\mathbf{u}_1^{w'} & \mathbf{u}_2^{w'}\end{bmatrix}$ of filtered effective speech subspace $U_j^{w'}$ for $1 \leq j \leq 2$. In this figure, $(\mathbf{v}_f \cdot \mathbf{u}_1^{w'})\mathbf{u}_1^{w'}$ and $(\mathbf{v}_f \cdot \mathbf{u}_2^{w'})\mathbf{u}_2^{w'}$ denote the orthogonal projec-

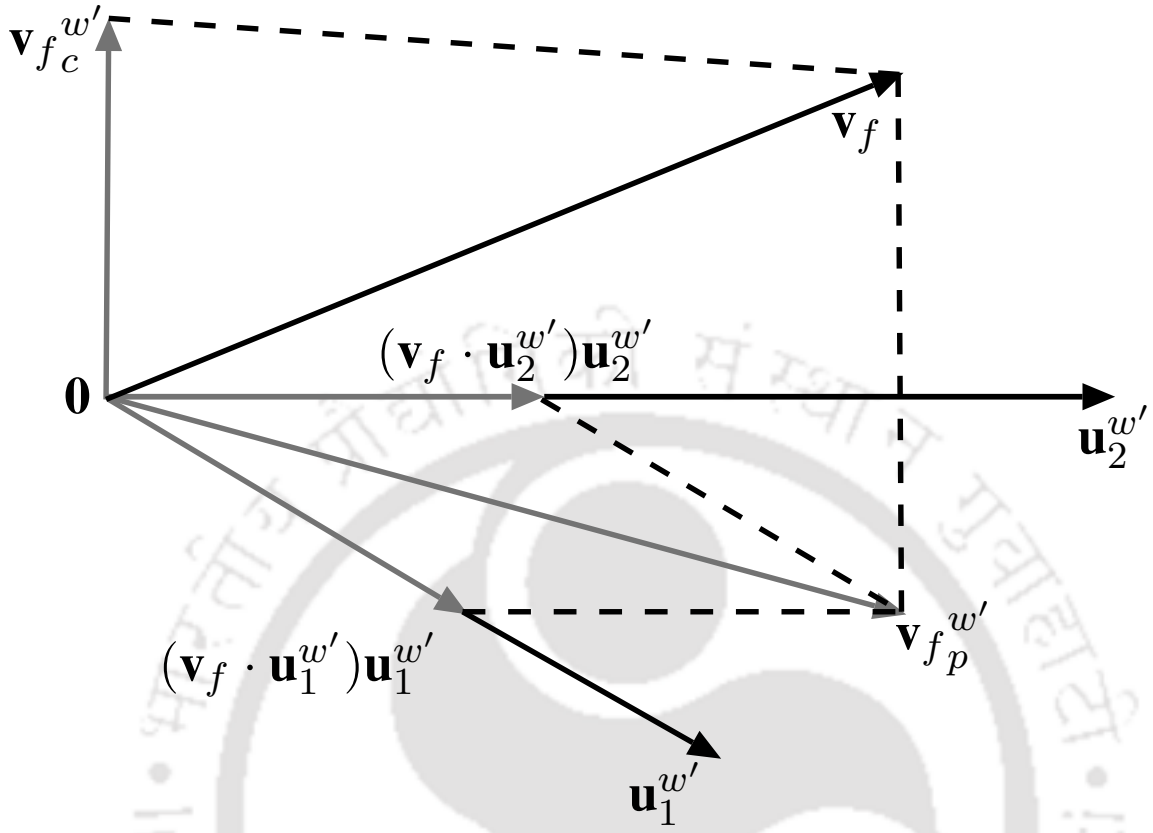


Figure 3.3: Orthogonal projection of stressed speech onto the filtered effective speech subspace.

tion component of stressed speech feature vector $\mathbf{v}_f$ onto the basis vectors $\mathbf{u}_1^{w'}$ and $\mathbf{u}_2^{w'}$, respectively. The normalized stressed speech feature vector $\mathbf{v}_{f_p}^{w'}$ is the summation of $(\mathbf{v}_f \cdot \mathbf{u}_1^{w'})\mathbf{u}_1^{w'}$ and $(\mathbf{v}_f \cdot \mathbf{u}_2^{w'})\mathbf{u}_2^{w'}$ as given in Eq. 3.6. The complement orthogonal projection component is represented by $\mathbf{v}_{f_c}^{w'}$. It is evident that, the normalization of stress information using the proposed approach depends on the number of basis vectors of filtered effective speech subspace $U_j^{w'}$. In this work, the proposed stress normalization algorithm has been evaluated by varying the number of basis vectors j, $1 \leq j \leq K$ to analyse the effectiveness of the number of basis vectors.

3.2 The Low-rank Subspace Projection Using Heteroscedastic Linear Discriminant Analysis

The subspace projection of features and model parameters onto the another common subspace, which consists of decorrelated characteristics leads to less number of parameters for the robust and the precise representation of speech patterns onto the feature- and the model-space, respec-

tively [148–150]. Therefore, after normalizing the stress information using the proposed orthogonal projection-based linear subspace modelling technique, the correlation values for the features and the model parameters are reduced by adapting the feature- and the model-space into the decorrelated subspace. To further increase the robustness and to reduce the computational complexity of the proposed stress normalization method, the rank of subspace projection matrix is reduced. The proposed subspace projection is accomplished in the maximum-likelihood linear transformation (MLLT)-based semi-tied adaptation framework [148] as described in block diagram shown in Figure 3.4. The linear transformation matrix for adaptation is learned on the training data using the heteroscedastic linear discriminant analysis (HLDA)-based low-rank subspace projection method [149, 150]. HLDA maps the K-dimensional subspace onto the common subspace (having dimension L, $L \leq K$), which consists of the property of minimum correlation between the features. It is hypothesized that, the L-dimensional subspace comprises the information of the distinctive mean and variance for each class, c ($1 \leq c \leq C$, C denotes the total number of class). It is also assumed that, the remaining $(K - L)$-dimensional subspace, which is called as the nuisance subspace contains the same mean and variance information of each class. The feature vector $\mathbf{x}$ having size $K \times 1$ of neutral or stressed speech data after performing the proposed subspace projection-based approach for stress normalization are linearly transformed using the $K \times K$ non-singular matrix $\mathbf{H}$ as follows

$$\mathbf{z} = \mathbf{H}\mathbf{x} = \begin{bmatrix} \mathbf{H}_L\mathbf{x} \\ \mathbf{H}_{K-L}\mathbf{x} \end{bmatrix} = \begin{bmatrix} \mathbf{z}_L \\ \mathbf{z}_{K-L} \end{bmatrix} \tag{3.7}$$

where, $\mathbf{H}_L$ and $\mathbf{H}_{K-L}$ contain the first L and the remaining $(K-L)$ rows of matrix $\mathbf{H}$, respectively. The vector $\mathbf{z}$ is the linearly transformed vector. The first L elements of vector $\mathbf{z}$ i.e. $\mathbf{z}_L$ have the properties of the minimum correlation between them. The remaining $(K - L)$ elements of vector $\mathbf{z}$ i.e. $\mathbf{z}_{K-L}$ represent the nuisance features. The vector $\mathbf{z}_L$ forms L dimensional decorrelated feature vector. The maximum-likelihood (ML) estimate of the k^{th} row $\mathbf{h}_k$ of $\mathbf{H}$ using the training data is given as,

$$\hat{\mathbf{h}}_k = \mathbf{c}_k \mathbf{G}_k{}^{-1} \sqrt{\frac{N}{\mathbf{c}_k \mathbf{G}_k{}^{-1} \mathbf{c}_k^T}} \tag{3.8}$$

where,

$$\mathbf{G}_k = \begin{cases} \sum_{c=1}^{C} \dfrac{N_c}{\mathbf{h}_k \mathbf{W}_c \mathbf{h}_k{}^T} \mathbf{W}_c & \text{if } 1 \leq k \leq L \\ \dfrac{N}{\mathbf{h}_k \mathbf{T} \mathbf{h}_k{}^T} \mathbf{T} & \text{if } L < k \leq K \end{cases}$$

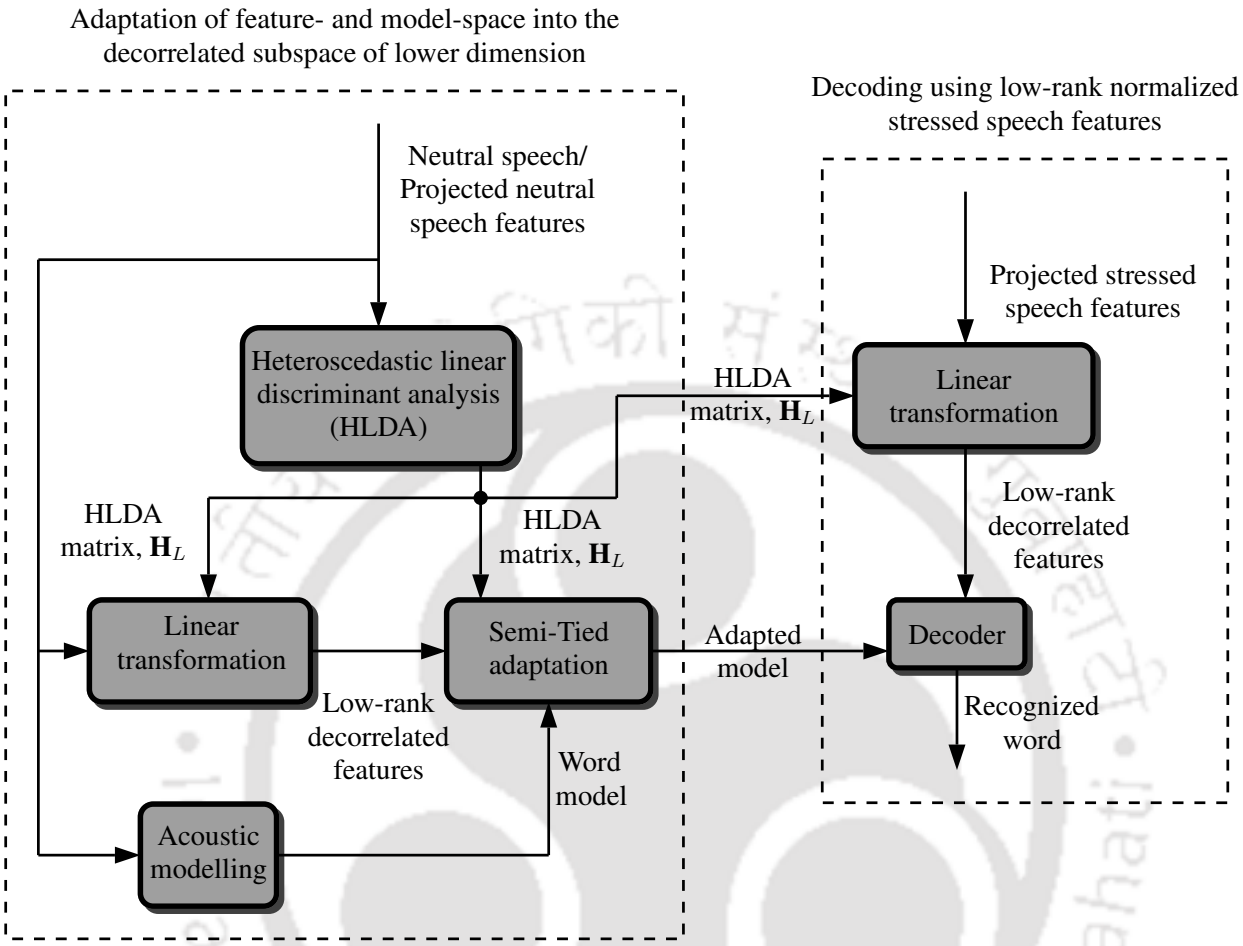


Figure 3.4: The adaptation of feature- and model-space onto a common decorrelated subspace for the recognition of stressed speech.

where, $\mathbf{c}_k$ denotes the k^{th} row of cofactor of the current estimate $\widehat{\mathbf{H}}$ of matrix $\mathbf{H}$. $\hat{\mathbf{h}}_k$ is the current estimate of the k^{th} row of $\widehat{\mathbf{H}}$. N_c represents the number of feature vectors, which belong to the class c of training data. The total number of feature vectors of the training data is denoted by $N = \sum_{c=1}^{C} N_c$. The matrix $\mathbf{W}_c$ represents the within class covariance matrix of the training data of c^{th} class. Whereas, the matrix $\mathbf{T}$ denotes the covariance matrix for the training data. The features of neutral and stressed speech utterances obtained after performing the proposed stress normalization methods are projected onto a common decorrelated subspace using the transformation matrix $\mathbf{H}_L$ as described in block diagram shown in Figure 3.4. The model-space is also adapted to this decorrelated subspace using the same transformation matrix $\mathbf{H}_L$, and the decorrelated features of the neutral speech data. The adapted acoustic models are decoded using the corresponding reduced dimensional decorre-

lated features of normalized stressed speech utterances. In this work, the dimension of this common subspace has been varied by varying the rank of transformation matrix $\mathbf{H}_L$ between 1 to K to observe the effectiveness of proposed stress normalization techniques onto the lower dimensional subspace.

3.3 Performance Evaluation of Linear Subspace Modeling

In this Section, the effectiveness of the proposed orthogonal projection-based linear subspace modelling technique for normalizing the stress-specific attributes is measured onto the two distinct frameworks namely: the stressed speech recognition and the error analysis, respectively. The task involving the speech recognition for the stressed speech are implemented using the automatic speech recognition (ASR) system. The performances of various ASR systems developed in this work are evaluated using the word error rate (WER) metric and is computed using Eq. 2.9 as given in Chapter 2. The error analysis is performed by exploring the Kullback Leibler (KL) divergence [140, 141] and the correlation coefficient [1, 3, 143, 147, 151, 152] metrics for measuring the similarities between the neutral and the stressed speech patterns. The illustrations highlight the description about the experimental setup and the performance evaluation of the proposed stress normalization method.

3.3.1 Experimental Setup

The performance of the proposed stress normalization method by exploiting the stressed speech recognition is evaluated using the Speech Under Simulated Stress Condition (SUSSC) database reported in [118] as described in Section 2.1 in Chapter 2. The acoustic model is trained using 2322 utterances of neutral speech. The test dataset consists of 700, 700, 594, 594 and 700 utterances of neutral, angry, sad, lombard and happy speech, respectively. All speech data used are digitized at a sampling frequency of 8 kHz with 16 bits/sample resolution and segmented using a Hamming window of length 20 msec, with a frame shift of 10 msec. Two kinds of speech features are explored for speech parametrization namely: 39-dimensional TEO-CB-Auto-Env features [44, 45] and 13-dimensional Mel-frequency cepstral coefficients (MFCCs) features [100]. First 12-dimensional base MFCCs ($C1$-$C12$) are determined using a 21-channel Mel-filterbank. The energy parameter is added as the zeroth coefficient making the base feature dimension equal to 13 ($C0 - C12$).

The error analysis using the Kullback Leibler (KL) divergence and the correlation coefficient metrics is performed over the same training and test datasets of neutral and stressed speech utterances

of SUSSC database as employed in stressed speech recognition cases, respectively. The speech signals are parameterized into the 39-dimensional TEO-CB-Auto-Env features and the 13-dimensional MFCCs ($C0 - C12$) features similar to as extracted for the tasks involving the speech recognition for the stressed speech.The KL divergence metric is evaluated over the Gaussian mixture models (GMMs) [112] comprising the mixture of 32 diagonal Gaussian densities.

The set of basis vectors of speech subspace for the accomplishment of proposed stress normalization technique are estimated using $R = 256$ numbers of clusters in k-means clustering method as discussed in Subsection 3.1.1. The 39-dimensional TEO-CB-Auto-Env and the 13-dimensional MFCC features create the matrix $\mathbf{C}^w$ having size 39×256 and 13×256, respectively. This, in turns, these two types of speech parameterization approaches have resulted in the matrix $\mathbf{U}^w$ of size $K \times K$ as 39×39 and 13×13, respectively. Therefore, using 39-dimensional TEO-CB-Auto-Env and 13-dimensional MFCC features, total 39 and 13 sets of basis vectors $\{\mathbf{U}_j^w\}_{j=1}^{39}$ and $\{\mathbf{U}_j^w\}_{j=1}^{13}$ are tested as the basis vectors for the speech subspaces, respectively.

In modelling paradigm, we have explored the speaker independent (SI) ASR system employing acoustic models based on continuous density left-to-right hidden Markov model (HMM), which is also referred to as the GMM-HMM system as discussed in Section 2.3 in Chapter 2. The GMM-HMM system comprises the 10 states, in which first and last states are non-emitting states and others are emitting states. The acoustic input in each state is modeled using the 3 diagonal covariance Gaussian density mixtures. All the GMM-HMM systems employed in this work are developed using the Hidden Markov Model Toolkit (HTK) [153].

3.3.2 Performance Evaluation

In this Subsection, the performances of the proposed linear subspace modelling approach based on orthogonal projection for stress normalization technique is evaluated using the aforementioned stressed speech recognition and error analysis approaches.

Stressed Speech Recognition: The recognition performances for the stressed speech using the proposed orthogonal projection-based linear subspace modelling technique over the TEO-CB-Auto-Env and the MFCC features are summarized in Table 3.1 and Table 3.2, respectively. The performances of speech recognition over all stress classes studied in this work are evaluated by varying the number of basis vectors j from $1 \leq j \leq 39$ and $1 \leq j \leq 13$ using the TEO-CB-Auto-Env and

Table 3.1: The recognition performances for stressed speech (WER in %) using the proposed stress normalization technique employing the orthogonal projection-based linear subspace modelling approach using the default 39-dimensional TEO-CB-Auto-Env features.

Stress Class	Conventional-case	Stress normalization using linear subspace modelling technique									
		Number of basis vectors corresponding to the highest eigenvalues, j									
		1	5	10	15	20	33	34	35	36	39
Angry	37.86	99.6	45.6	41.3	41	40.7	38.1	37.3	**37.1**	37.6	37.9
Sad	30.81	97.1	39.6	34.7	34.3	31.3	30.8	30.5	29.8	**29.5**	30.8
Lombard	26.26	99.7	38.7	31.3	29.3	28.8	**24.6**	24.9	25.1	25.4	26.3
Happy	21.29	98.4	30.3	22.7	21.4	20.4	20.4	20.1	**19.6**	20.3	21.3
Average	29.05	98.7	38.5	32.5	31.5	30.3	28.5	28.2	**27.9**	28.2	29.1

the MFCC features, respectively. The effectiveness of the proposed method is quantified by comparing with the performances of baseline ASR system, which is also referred to as the conventional cases. The baseline ASR systems are developed on the 39-dimensional TEO-CB-Auto-Env and 13-dimensional MFCC features of original neutral speech utterances of training database. In this work, the default dimensions are consider as the $L = 39$ and the $L = 13$ using the 39-dimensional TEO-CB-Auto-Env and the 13-dimensional MFCC features, respectively. The WER values summarized for the conventional cases correspond to the recognition performances using the baseline ASR system for the recognition of original stressed speech utterances without implementing the proposed stress normalization method. In these tables, the bold face WER values are representing the best case performances. The WERs reported in Table 3.1 illustrate that, the best recognition performances with the maximum relative decrement in the WER values over the conventional cases are achieved, when 35, 36, 33 and 35 eigenvectors corresponding to the highest eigenvalues are used to filter an effective speech subspace under angry, sad, lombard and happy conditions, respectively. Using these specific sets of basis vectors, the maximum degradation in WER of 2%, 4.25%, 6.32% and 7.94% are obtained, when compared to the conventional results for the recognition of angry, sad, lombard and happy speech, respectively. The WERs depicted in Table 3.2 show that, using MFCC

Table 3.2: The recognition performances for stressed speech (WER in %) using the proposed stress normalization technique employing the orthogonal projection-based linear subspace modelling approach using the default 13-dimensional MFCC features.

Stress Class	Conven-tional case	Stress normalization using linear subspace modelling technique									
		Number of basis vectors corresponding to the highest eigenvalues, j									
		1	3	5	7	8	9	10	11	12	13
Angry	51.3	95.4	69	62.9	55.4	56.3	54.9	53	51.4	52.7	**51.3**
Sad	36.9	96.1	65.8	52.4	46	39.9	39.6	38.4	36.7	**36.4**	36.9
Lombard	33.8	95.6	63.5	51	42.4	41.1	39.9	38.4	35.5	36	**33.8**
Happy	23.4	95.9	51.9	36.7	29.1	28.3	26.3	25.6	24.4	**23.1**	23.4
Average	36.3	95.7	62.5	50.7	43.2	41.4	40.2	38.8	37	37	**36.3**

features, when 13, 12, 13 and 12 eigenvectors corresponding to the highest eigenvalues are used as the basis vectors for the speech subspace, the maximum degradation in the WERs are obtained in comparison to the conventional cases for the recognition of angry, sad, lombard and happy speech, respectively. The improved performances of stressed speech recognition using TEO-CB-Auto-Env features demonstrate that, the orthogonal projection onto these specific sets of basis vectors i.e. $\{\mathbf{u}_j^{w'}\}_{j=1}^{35}$, $\{\mathbf{u}_j^{w'}\}_{j=1}^{36}$, $\{\mathbf{u}_j^{w'}\}_{j=1}^{33}$ and $\{\mathbf{u}_j^{w'}\}_{j=1}^{35}$ helps in normalizing the stress-specific attributes of angry, sad, lombard and happy speech, respectively. Similarly, using MFCC features, the specific set of basis vectors, $\{\mathbf{u}_j^{w'}\}_{j=1}^{12}$ normalizes the stress information of both the sad and the happy speech. These specific sets of basis vectors effectively span the speech-specific attributes. Consequently, the orthogonal projection of stressed speech onto the speech subspaces spanned by these specific sets of basis vectors normalize the stress information. The resulting orthogonally projected features of stressed speech model the linear subspace comprising the features having normalize stress information. This linear subspace represents the linear characteristics between the speech- and the stress-specific attributes. The proposed linear subspace modelling technique effectively reduces the acoustic mismatches between the different speech units of neutral and stressed speech.

After normalizing the stress-specific attributes using the specific sets of basis vectors, the feature-

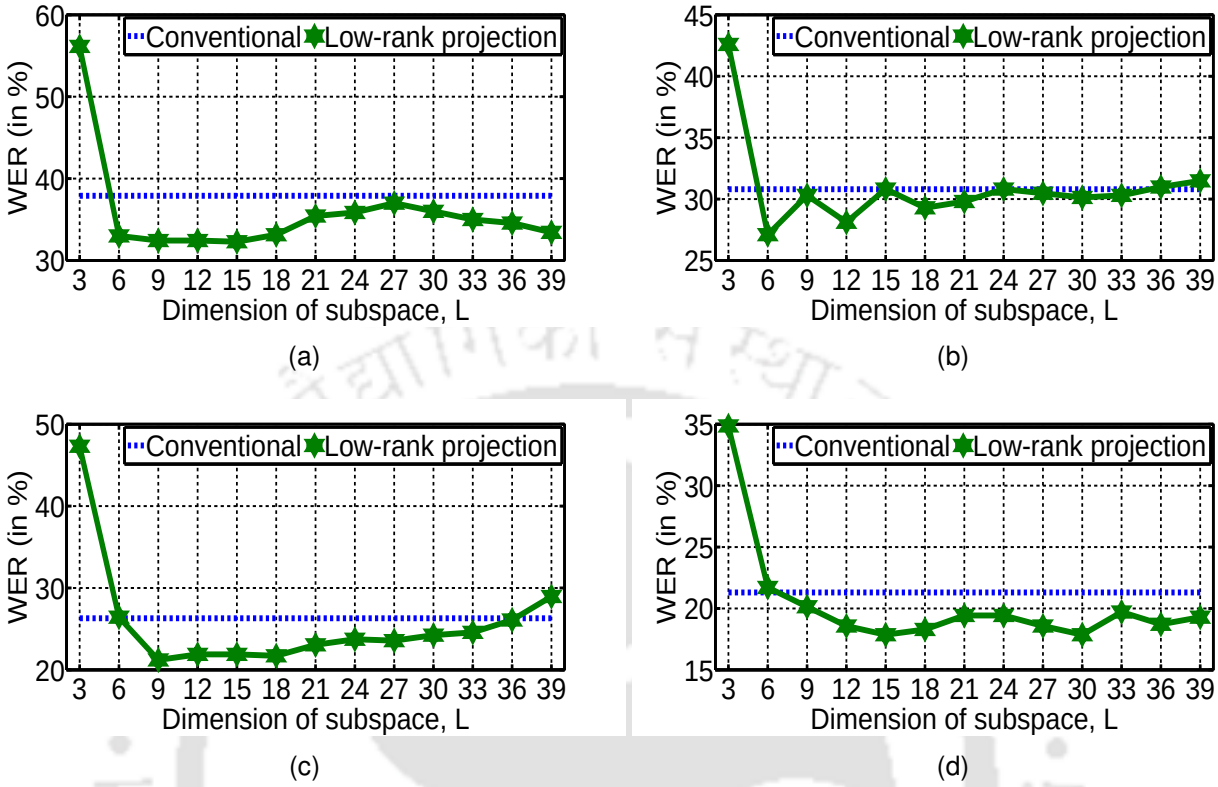


Figure 3.5: Change in the WERs using the proposed stress normalization technique employing the orthogonal projection-based linear subspace modelling followed by the HLDA-based low-rank subspace projection method using the TEO-CB-Auto-Env features. (a), (b), (c) and (d) represent the stressed speech recognition performances under angry, sad, lombard and happy conditions, respectively.

and the model-space are adapted onto the decorrelated subspace having dimension lower than that of the default dimension. The proposed adaptation technique has exploited the HLDA-based low-rank subspace projection in the MLLT semi-tied adaptation approach. As discussed earlier, $L = 39$ and $L = 13$ are chosen as the default dimension corresponding to the 39-dimensional TEO-CB-Auto-Env and the 13-dimensional MFCC features, respectively. The performances of low-rank subspace projection using TEO-CB-Auto-Env and MFCC features over all the explored stress classes are depicted by the WER-profiles shown in Figure 3.5 and Figure 3.6, respectively. The conventional cases given in these figures corresponds to the case of using default dimensional features of original neutral and stressed speech utterances. The rank of the subspace projection matrix is varied from 3 to 39 and 3 to 13 using the TEO-CB-Auto-Env and the MFCC features in steps of 3 and 1, to obtain the WER-profiles for low-rank subspace projection cases, respectively. For every reduction in the rank of the

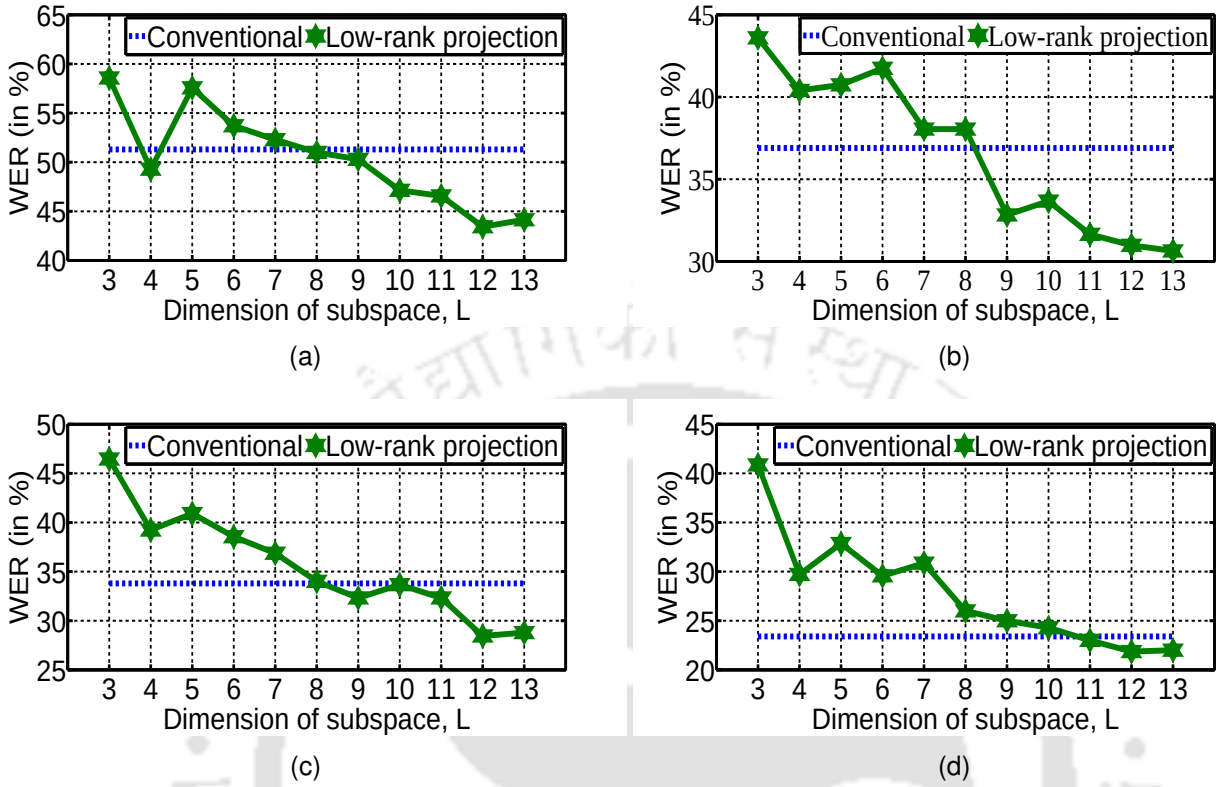


Figure 3.6: Change in the WERs using the proposed stress normalization technique employing the orthogonal projection-based linear subspace modelling followed by the HLDA-based low-rank subspace projection method using the MFCC features. (a), (b), (c) and (d) represent the stressed speech recognition performances under angry, sad, lombard and happy conditions, respectively.

projection matrix, a separate ASR system is trained and tested with corresponding reduced dimensional features as described in block diagram shown in Figure 3.4. From the reported WER-profiles, it is evident that, the projection of neutral and normalized stressed speech onto a decorrelated subspace having dimension lower than that of the default dimension has resulted in reduction of WER values. This behavior is consistently exhibited for all types of features explored in this work. When the normalized TEO-CB-Auto-Env features of angry, sad, lombard and happy speech are linearly transformed onto the $L = 15$, $L = 6$, $L = 9$ and $L = 15$ dimensional decorrelated feature subspaces, the maximum improved performances of speech recognition are obtained in comparison to the conventional results. On comparing the best performing low-rank features out of the default rank features, the maximum degradation of 14.8%, 12%, 19.4% and 16% in WER values are noted for the recognition of angry, sad, lombard and happy speech as depicted in Figure 3.5(a)$-$Figure 3.5(d), respectively. Whereas, using MFCC features, $L = 12$, $L = 13$, $L = 12$ and $L = 12$ dimension of

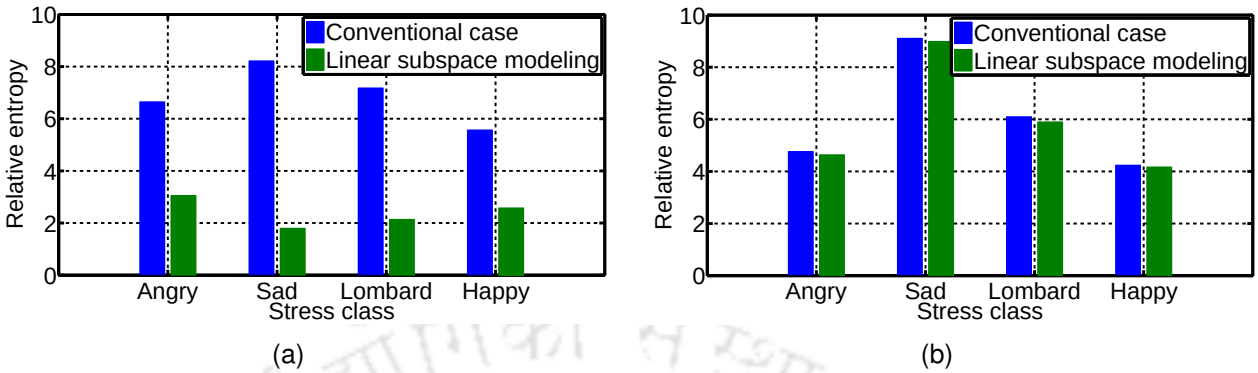
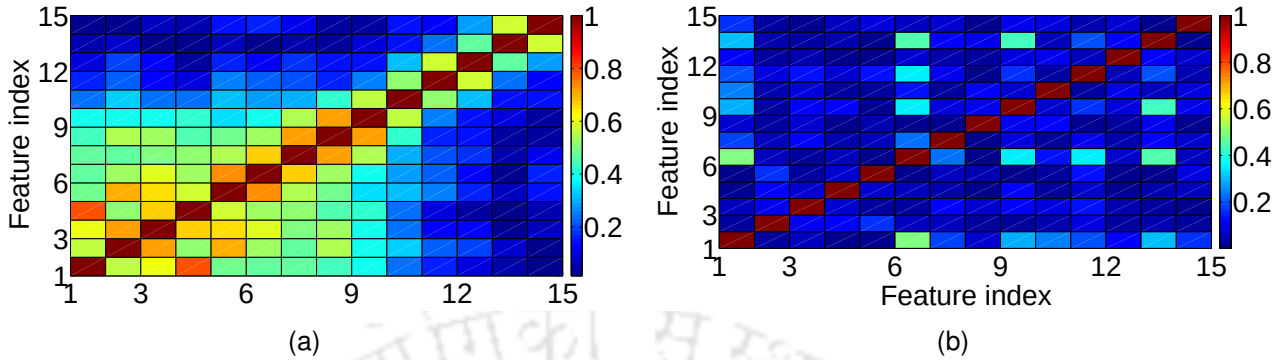


Figure 3.7: The error analysis by exploring the KL divergence metric using the proposed stress normalization technique employing the orthogonal projection-based linear subspace modelling followed by the HLDA-based low-rank subspace projection method. (a) and (b) represent the KL divergence values for the TEO-CB-Auto-Env and the MFCC features, respectively.

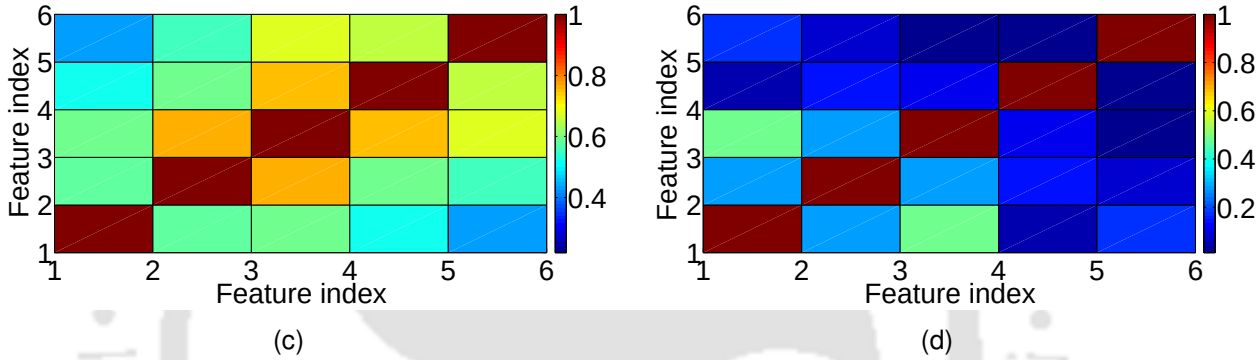
decorrelated features of normalized angry, sad, lombard and happy speech lead to the maximum relative improvement in the performance of speech recognition, respectively. The performances of stressed speech recognition are improved by a relative decrement in the WER of 15.4%, 17.1%, 16% and 6.4% over the conventional cases under angry, sad, lombard and happy conditions as shown in Figure 3.6(a)−Figure 3.6(d), respectively. The improved performances of speech recognition over all the studied stress classes and feature types with reduced values of WER signify the reduction in the variance mismatches between the different speech units of neutral and stressed speech. The proposed low-rank subspace projection helps in retaining the speech-specific attributes in the low frequency bandwidth of both the neutral and the stressed speech. Consequently, the lower dimensional features of neutral and stressed speech create the similar characteristics for the training and the test environments of ASR system. This, in turns, the ASR system developed on the lower dimensional features of neutral speech has resulted in the improved speech recognition performances for the stressed speech in comparison to the performances of stressed speech recognition obtained using the ASR system trained on the default dimensional features of neutral speech utterances.

Error Analysis: The performance of the proposed orthogonal projection-based linear subspace modelling technique is further quantified by exploiting the error analysis using the SUSSC database reported in [118]. The error analysis is measured by evaluating the similarity between the Gaussian-subspaces developed using the neutral speech (training data) and the test speech utterances pro-

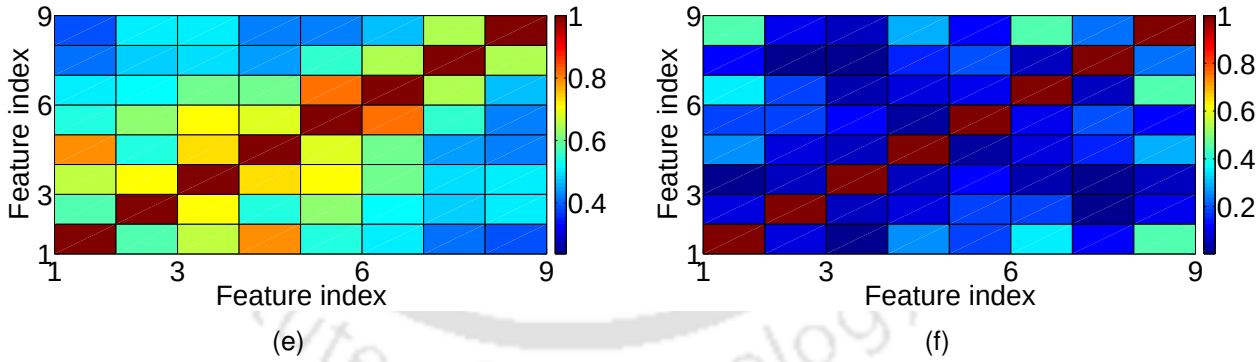
duced under four stress conditions namely: angry, sad, lombard and happy, respectively. The similarity between the Gaussian-subspaces are evaluated using the Kullback Leibler (KL) divergence metric or relative entropy between the Gaussian mixture models (GMMs) consisting of mixture of 32 diagonal Gaussian densities [140, 141]. The KL divergence values are determined over both the 39-dimensional TEO-CB-Auto-Env and the 13-dimensional base MFCCs ($C0 - C12$) features as shown in bar plots in Figure 3.7. In these bar plots, the KL divergence values for the conventional case are evaluated by determining the relative entropy between the GMMs trained on the aforementioned features of original neutral and stressed speech signals. The KL divergence values for the orthogonal projection case corresponds to the relative entropy between the GMMs trained over the specific lower dimensional decorrelated features of neutral and normalized stressed speech determined using the proposed orthogonal projection-based linear subspace modelling with HLDA-based low-rank subspace projection technique. The relative entropy reported in these plots show that, the subspace projection of neutral and normalized stressed speech features onto a common specific lower dimensional decorrelated subspace reduces the KL divergences between the GMMs in comparison to the KL divergence values obtained in the conventional cases. The large scale of relative decrement in the KL divergence values are noted using the TEO-CB-Auto-Env features in comparison to the MFCC features as depicted in bar plots shown in Figure 3.7. Using proposed method, the KL divergences between the GMMs trained over TEO-CB-Auto-Env features are decreased by 54.13% (6.65 to 3.05), 78.10% (8.22 to 1.8), 70.19% (7.18 to 2.14) and 53.68% (5.57 to 2.58), when compared to the KL divergences determined in the conventional cases under angry, sad, lombard and happy conditions, respectively. Whereas, the relative decrement in the KL divergence values by 2.52% (4.76 to 4.64), 1.32% (9.11 to 8.99), 3.28% (6.10 to 5.90) and 1.68% (4.24 to 4.17) are noted between the GMMs learned over the MFCC features in comparison to the conventional results for angry, sad, lombard and happy speech, respectively. The reduced values of KL divergence revealed that, the Gaussian-subspaces corresponding to the neutral and the normalized stressed speech of specific lower dimension exhibit the similar characteristics. The similar properties between the Gaussian-subspaces demonstrate that, the proposed orthogonal projection of stressed speech onto the filtered effective speech subspace followed by the low-rank subspace projection effectively reduces the discrepancy between the spectral distribution of the different speech units of the neutral and the normalized stressed speech.



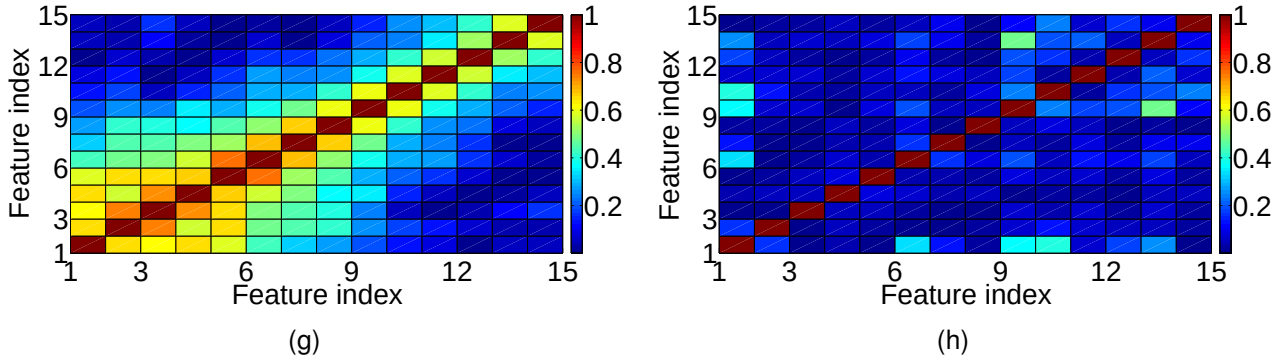
Correlation coefficient matrix for $L = 15$ dimensional features; (a) angry speech and (b) normalized angry speech



Correlation coefficient matrix for $L = 6$ dimensional features; (c) sad speech and (d) normalized sad speech

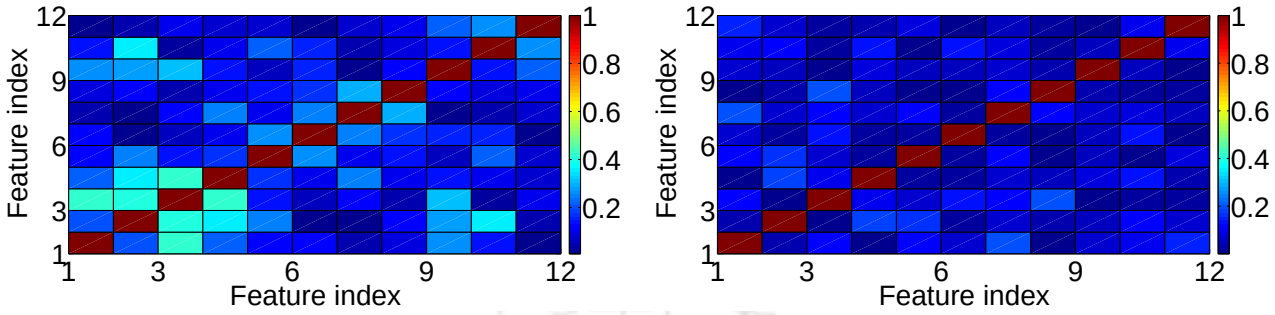


Correlation coefficient matrix for $L = 9$ dimensional features; (e) lombard speech and (f) normalized lombard speech

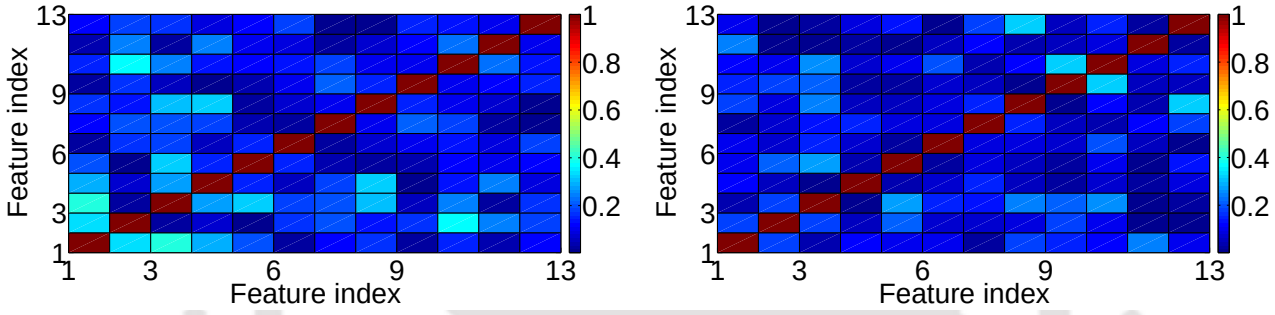


Correlation coefficient matrix for $L = 15$ dimensional features; (g) happy speech and (h) normalized happy speech

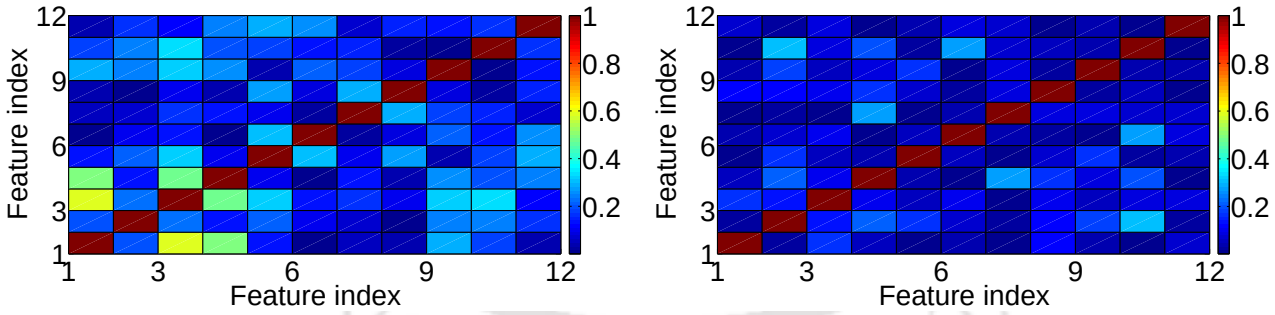
Figure 3.8: The correlation between the specific lower dimensional TEO-CB-Auto-Env features of original stressed speech and normalized stressed speech derived using the proposed linear subspace modelling technique along with the HLDA-based low-rank subspace projection method.



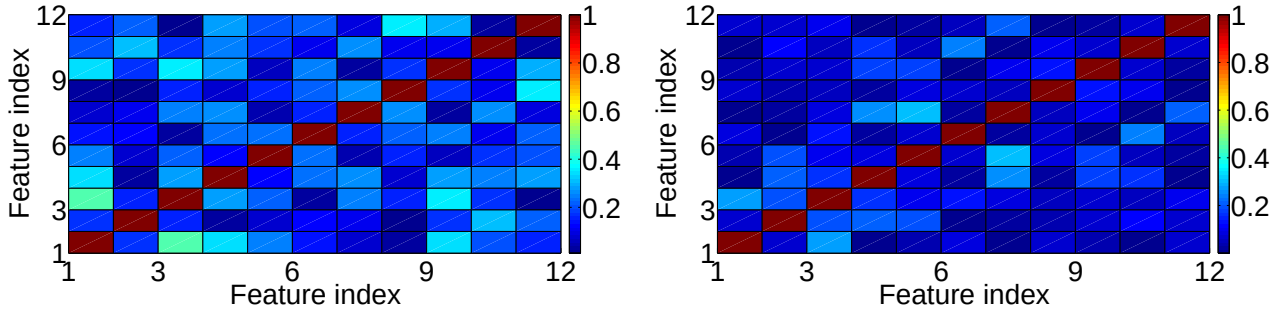
Correlation coefficient matrix for $L = 12$ dimensional features; (a) angry speech and (b) normalized angry speech



Correlation coefficient matrix for $L = 13$ dimensional features; (c) sad speech and (d) normalized sad speech



Correlation coefficient matrix for $L = 12$ dimensional features; (e) lombard speech and (f) normalized lombard speech



Correlation coefficient matrix for $L = 12$ dimensional features; (g) happy speech and (h) normalized happy speech

Figure 3.9: The correlation between the specific lower dimensional MFCC features of original stressed speech and normalized stressed speech derived using the proposed linear subspace modelling technique along with the HLDA-based low-rank subspace projection method.

The WER-profiles summarized in Figure 3.5 and Figure 3.6 illustrate that, the adaptation of feature- and model-space onto the common decorrelated subspaces having specific lower dimension reduces the variance mismatch between the training and test environments of ASR system. The resulting ASR systems have resulted in the improved stressed speech recognition performances, when compared to the conventional cases. The relative decrement in the WERs demonstrate that, the proposed low-rank subspace projection reduces the discrepancy between the spectral distribution of the different speech units of neutral and normalized stressed speech. Moreover, the deteriorated values of relative entropy between the Gaussian-subspaces, learned over the specific lower dimensional decorrelated features of neutral and normalized stressed speech in comparison to the conventional cases have again manifested the reduction in the acoustic dissimilarities as shown in Figure 3.7. The adaptation of feature- and model-space onto the specific lower dimensional decorrelated subspace has resulted in the robust and the precise representation of stressed speech with illuminating patterns onto the feature-space and the less number of informative model parameters onto the model-space. In this work, the degree of decorrelation between the features is studied by plotting the correlation coefficient matrix. Figure 3.8 and Figure 3.9 summarize the correlation between the decorrelated TEO-CB-Auto-Env and MFCC features of normalized stressed speech derived using the proposed HLDA-based low-rank subspace projection onto the specific lower dimensional decorrelated subspace. In these figures, correlation coefficient matrices are also plotted for the specific lower dimensional features of original stressed speech utterances and are referred to as the conventional cases. The reduced dimension of original stressed speech features are determined by selecting only the first that specific number elements of feature vectors. The correlation coefficient matrices plotted in Figure 3.8 demonstrate that, when TEO-CB-Auto-Env features of normalized angry, sad, lombard and happy speech are linearly transformed onto the decorrelated feature subspace of specific lower dimension $L = 15$, $L = 6$, $L = 9$ and $L = 15$, the average values of correlation coefficients are decreased in a large scale, when compared to the average value of correlation coefficients determined in the conventional cases, respectively. The average value of correlation coefficients are reduced by 56.84% (0.35 to 0.15), 55.49% (0.62 to 0.28), 60.88% (0.57 to 0.22) and 57.11% (0.34 to 0.15) in comparison to the correlation coefficients obtained in conventional cases under angry, sad, lombard and happy conditions, respectively. Using MFCC features, when normalized angry, sad, lombard and happy speech are linearly transformed into the decorrelated feature subspace of specific lower

dimension $L = 12$, $L = 13$, $L = 12$ and $L = 12$, the average values of correlation coefficients are decreased significantly over the average values of conventional correlation coefficients, respectively, as shown in Figure 3.9. The average values of correlation coefficients are decreased by 37.08% (0.23 to 0.15), 16.20% (0.20 to 0.17), 33.57% (0.23 to 0.15) and 33.74% (0.24 to 0.16) in comparison to the average value of correlation coefficients determined using the conventional cases for angry, sad, lombard and happy speech, respectively. These experimental results demonstrate that, the HLDA-based low-rank subspace projection in MLLT-based semi-tied adaptation technique is very effective for the subspace projection of features and mode parameters onto the specific lower dimensional decorrelated subspace. The reduced value of correlation for the neutral speech features, normalized stressed speech features and the model parameters have resulted in the effective representation onto the feature and the model-space, respectively. The proposed linear subspace modelling technique by exploring the orthogonal projection approaches along with the HLDA-based low-rank subspace projection technique have significantly reduced the acoustic mismatch between the neutral and the stressed speech and developed the effective stress normalization method.

3.4 Non-linear Subspace Modeling Using Polynomial Function

The proposed stress normalization method using the linear subspace modelling approach by exploring the orthogonal projection technique is found to be effective for normalizing the stress information as described in Section 3.3. Although stress normalization using this method with promising results are reported, there are still a slight improvement in the recognition performances of speech under some stress conditions, when compared to the conventional cases. The proposed method using TEO-CB-Auto-Env features has resulted in the relative decrement in the WER of 2% (37.86% to 37.1%) in comparison to the conventional case for the recognition of angry speech as shown in Table 3.1. In other cases, the WER values depicted in Table 3.2 illustrate that, the assumption of linear relationship between the speech and the stress information using the MFCC features leads to the recognition performances of angry and lombard speech are identical to as obtained from the conventional ASR systems. These experimental observations demonstrate that, the modelling of linear subspace has appeared promising for normalizing the stress information in specific stress classes. Stress introduces the non-linearity in the pattern of stressed speech in comparison to the pattern of neutral speech. In literature, numerous studies have been manifested the non-linear characteristics

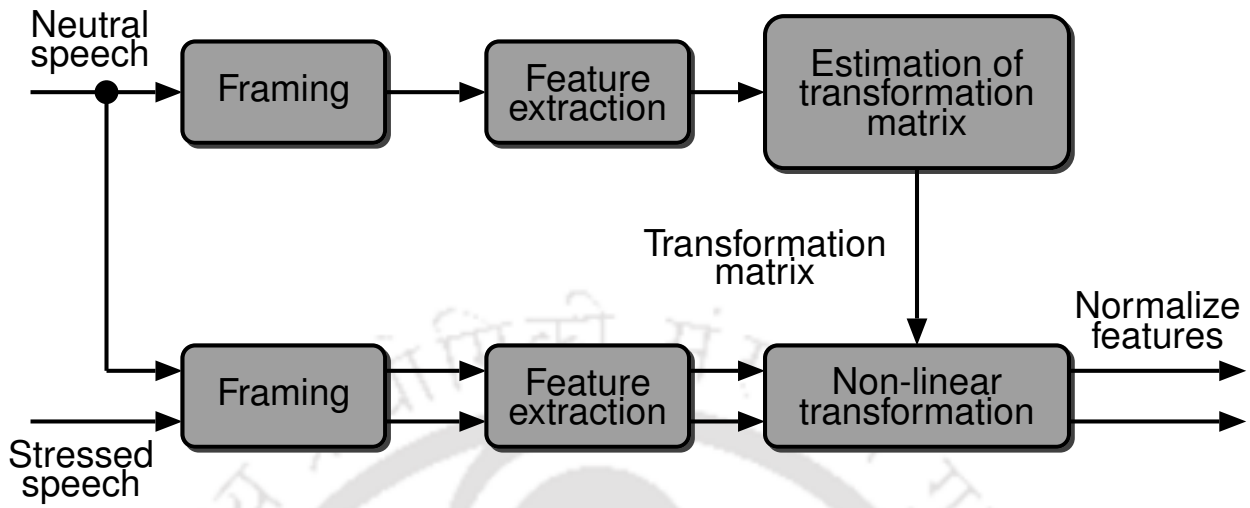


Figure 3.10: The proposed non-linear subspace modelling technique for stress normalization.

of speech production system under stress condition [6, 22, 33, 41–45]. Therefore, the analysis of stressed speech onto the non-linear subspace may help in developing the robust and the computationally efficient stress normalization algorithms.

In this Section, the stress normalization technique is designed by studying the variance mismatch between the neutral and the stressed speech onto the non-linear subspace. To reduce the acoustic variabilities, both the neutral and the stressed speech are projected onto a common subspace. The set of bases for this common subspace is estimated using the neutral speech data as it comprises the speech-specific attributes and it helps in spanning the speech information. This common subspace is generally referred to as the speech subspace. Consequently, the subspace projection of neutral and stressed speech onto the speech subspace can alleviate the effect of stress as well as retain the speech information. The proposed approach has explored the subspace projection through the non-linear transformation technique to model the non-linear subspace. In this work, the non-linear transformation has been accomplished onto the framework of polynomial function. The non-linearity between the speech- and the stress-specific attributes is investigated onto this non-linear subspace and it helps in separating them. The details involved in the proposed stress normalization algorithm is depicted in the block diagram shown in Figure 3.10. From this figure, it is evident that, the efficacy of non-linear subspace modelling depends on the estimation of an effective transformation matrix. The estimation of transformation matrix, which constitutes the acoustic properties of neutral speech data

can help in subspace projection of neutral and stressed speech onto the speech subspace. In addition to this, the effectiveness of the proposed non-linear transformation is measured by comparing it with the linear transformation-based subspace projection approach. In the following Subsections, we have described the methodology of the non-linear transformation to develop the non-linear subspace for normalizing the stress information. This is followed by the learning mechanism of transformation matrix. A brief discussion on the low-rank subspace projection method is also presented.

3.4.1 Non-linear Transformation onto the Speech Subspace

The primary objective of modelling of non-linear subspace is to separate the speech- and the stress-specific attributes. The non-linear subspace is modeled by exploiting the non-linear transformation technique. The feature vector **x** of speech signal (neutral or stressed speech) are projected onto the speech subspace through the non-linear transformation [142] as follows

$$\mathbf{y} = \mathbf{P}f(\mathbf{x}) \tag{3.9}$$

where, **y** represents the non-linearly transformed feature vector corresponding to the neutral/ stressed speech utterances onto the speech subspace using the transformation matrix **P**. The non-linear transformation is derived using the polynomial function $f(.)$ [142] given as

$$f(\mathbf{x}) = \sum_{p_n=0}^{P_n} \mathbf{x}^{p_n} \tag{3.10}$$

The order p_n of polynomial function is varied from $0 \leq p_n \leq P_n$. Using this polynomial function, the non-linearly transformed feature vector **y** can be expressed as

$$\begin{aligned}\mathbf{y} &= \mathbf{P}f(\mathbf{x}) \\ &= \mathbf{P}\sum_{p_n=0}^{P_n} \mathbf{x}^{p_n} \\ &= \mathbf{Pi} + \mathbf{Px} + \mathbf{Px}^2 + \mathbf{Px}^3 + \cdots + \mathbf{Px}^{P_n}\end{aligned} \tag{3.11}$$

where, the vector **i** is referred to as the identity vector. The subspace projection of feature vector **x** through the polynomial function yields the features constituting the summation of bases of speech subspace **Pi**, the linearly transformed features **Px**, the linearly transformed features i.e., $\mathbf{Px}^2$, ..., $\mathbf{Px}^{P_n}$ corresponding to the feature vector of higher order ($\mathbf{x}^2$, ..., $\mathbf{x}^{P_n}$) through the same transfor-

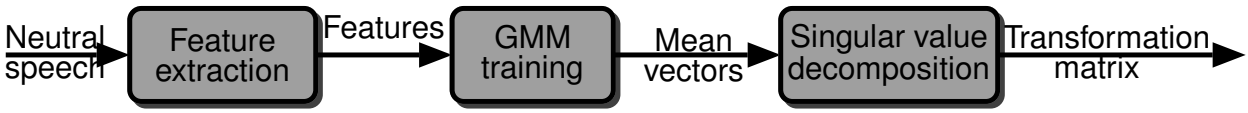


Figure 3.11: The proposed learning mechanism of transformation matrix.

mation matrix **P**. The non-linearly transformed feature vectors corresponding to the neutral and the stressed speech utterances are considered as the features having normalize stress information and constituting the speech-specific attributes. These non-linearly transformed features carry the similar acoustic characteristics and develop a non-linear subspace. As illustrated in Eq. (3.11), the efficacy of non-linear subspace modelling technique depends on the creation of transformation matrix **P**. In the following Subsection, we have described the learning mechanism of transformation matrix.

3.4.2 Learning Mechanism of Transformation Matrix

The proposed stress normalization method employs the subspace projection of neutral and stressed speech onto a common subspace using the non-linear transformation to reduce the acoustic mismatch between them. The common subspace, which comprises the speech-specific attributes, called as the speech subspace can assess the normalization of stress information with preserving the speech information. As discussed earlier, both the neutral and the stressed speech constitute the speech information. Therefore, the subspace projection of neutral and stressed speech through the transformation matrix, whose columns comprise the acoustic properties of neutral speech will help in projecting them onto the speech subspace. In order to address this, the transformation matrix is learned using neutral speech utterances. The learning mechanism incorporates the singular value decomposition (SVD) [142] on the mean parameter of Gaussian mixture model (GMM) [112] trained using the neutral speech as depicted in block diagram shown in Figure 3.11. The columns of transformation matrix **P** are considered as the bases of speech subspace and are estimated as follows.

Step I: A GMM, as the weighted sum of G-mixture Gaussian densities is trained using the features of neutral speech utterances (training database) belong to word w. Where w, $(1 \leq w \leq W)$ is the index for the number of words present in the training database. The set of mean

vectors $\mathbf{M}^w = \begin{bmatrix} \mathbf{m}_1^w & \mathbf{m}_2^w & .... & \mathbf{m}_G^w \end{bmatrix}$ corresponding to Gaussian density for each word captures the acoustic properties of the neutral speech and are arranged in $\mathbf{M} = \begin{bmatrix} \mathbf{M}^1 & \mathbf{M}^2 & .... & \mathbf{M}^W \end{bmatrix}$. The size of matrix **M** becomes $K \times WG$, where K is the dimension of feature vector.

Step II: In the next step, the SVD is performed over the matrix **M** to achieve the orthogonal characteristics as follows,

$$\mathbf{M} = \mathbf{USV}^{'} \tag{3.12}$$

where, **U** and **V** are orthonormal matrices having size $K \times K$ and $WG \times WG$, respectively. **S** is a $K \times WG$ diagonal matrix. The column vectors of **U** span the column space of **M** [142]. They are a set of independent vectors since they are mutually orthogonal [142]. These orthonormal vectors in **U** are utilized as the columns of the transformation matrix **P**.

Using the estimated transformation matrix **P**, both the neutral and the stressed speech are projected onto the speech subspace through the non-linear transformation by exploiting the polynomial function as the non-linear function. The resulting non-linearly transformed features of neutral and stressed speech create the non-linear subspace and are considered as the features constituting the similar acoustic properties. In addition to this, the subspace projection is also derived using the linear transformation through the same transformation matrix **P**. In this work, the proposed stress normalization algorithm employing the non-linear subspace modelling approach is contrasted with the stress normalization technique accomplished using the linear transformation method.

3.4.3 Decorrelation of Feature- and Model-Space

The adaptation of feature- and model-space onto the another common subspace constituting the decorrelated characteristics and dimension lower than that of the default dimension is reported to be very effective in enhancing the robustness of the proposed stress normalization method using the linear subspace modelling technique as discussed in Section 3.3. The subspace projection of neutral and normalized stressed speech features onto the specific lower dimensional decorrelated subspace has resulted in the significant improvement in the performances of stressed speech recognition, when compared to the conventional cases over all the studied features and stress classes as summarized in Figure 3.5 and Figure 3.6. Moreover, the relative decrement in the KL divergence values between the Gaussian-subspaces, developed using the specific lower dimensional decorrelated features of

neutral and normalized stressed speech have also demonstrated the acoustic similarities as shown in Figure 3.7. These observations motivate us to investigate the effectiveness of the proposed non-linear subspace modelling technique into the lower dimensional subspace. In order to address this, we have further explored the low-rank subspace projection using the heteroscedastic linear discriminant analysis (HLDA) [149, 150] in maximum-likelihood linear transformation (MLLT)-based semi-tied adaptation technique similar to as described in Section 3.2. To adapt the feature-space, the non-linearly transformed features corresponding to the neutral and the stressed speech are projected onto the lower dimensional decorrelated subspace using the transformation matrix $\mathbf{H}_L$ as shown in Figure 3.4. The model-space is adapted using the same transformation matrix $\mathbf{H}_L$ and the non-linearly transformed features of neutral speech utterances. For decoding, the adapted acoustic models are tested with correspond reduced dimensional decorrelated features of normalized stressed speech.

3.5 Performance Evaluation of Non-Linear Subspace Modeling

This Section addresses the performance evaluation of the proposed stress normalization algorithm developed by exploring the non-linear subspace modelling technique onto the framework of polynomial function. The effectiveness of the proposed stress normalization method is quantified by comparing it with the conventional and the linear transformation cases. The conventional case comprises the baseline automatic speech recognition (ASR) system trained on the features of original neutral speech data and tested using the original stressed speech utterances without implementing the proposed stress normalization method as discussed in Section 3.3. The efficacy of the proposed method is evaluated using the tasks involving the stressed speech recognition and the error analysis. The accuracies of all the ASR systems developed in this work are measured using the word error rate (WER) metric as given in Eq. 2.9. The error analysis is evaluated by exploiting the Kullback Leibler (KL) divergence and the correlation coefficient metrics. In the following paragraphs, the experimental setup and the performance evaluation of the proposed stress normalization method are described.

Experimental Setup: The performance of the proposed stress normalization method is evaluated using the experimental setup similar to as described in Subection 3.3.1. The experimental results obtained using the orthogonal projection-based linear subspace modelling technique have been reported to be better performances using the TEO-CB-Auto-Env features, when compared to the those

Table 3.3: The recognition performances for stressed speech (WER in %) using the proposed stress normalization technique employing the linear and non-linear transformation-based subspace projection method using the default 39-dimensional TEO-CB-Auto-Env features.

Stress class	Conven--tional case	Linear transfor--mation	Non-linear transformation									
			Order of polynomial function, p_n									
			1	2	3	4	5	6	7	8	9	10
Angry	37.86	31.57	31.14	29.86	29.43	29.14	28.29	31.14	29.43	**27.71**	29.43	29.29
Sad	30.81	26.26	**25.93**	27.95	27.95	28.96	28.45	28.96	29.12	28.79	30.47	28.79
Lombard	26.26	22.22	**20.03**	23.23	23.06	23.57	24.07	22.05	24.75	23.06	23.74	23.57
Happy	21.29	18.14	18.71	17.29	17.14	**16.29**	17.29	18.71	17	17.43	18.43	18.43
Average	29.05	24.55	**23.95**	24.58	24.47	24.49	24.54	25.21	25.07	24.25	25.52	25.02

using the MFCC features as summarized in Subection 3.3.2. Therefore, the non-linear subspace modelling approach for normalizing the stress-specific attributes has been evaluated using the TEO-CB-Auto-Env features. The transformation matrix **P** is learned using the GMM constituting $G = 32$ mixture of Gaussian densities developed on the neutral speech utterances of training database.

Stressed Speech Recognition: The recognition performances for the stressed speech using the proposed transformation-based subspace modelling approaches onto the default 39-dimensional speech subspace are summarized in Table 3.3. The bold face WER values depicted in this table correspond to the best case performances. A severe degradation in the performances are noted for the stressed speech cases, when compared to the recognition performance for the neutral speech. The WERs reported in this table show that, the non-linear transformation of both the neutral and the stressed speech features onto the speech subspace using the polynomial function reduces the acoustic dissimilarities between them more pronouncedly in comparison to the subspace projection derived using the linear transformation technique. The linear transformation of neutral and stressed speech onto the speech subspace have resulted in the relative improvement in the recognition per-

formances by the decrement in the WER values of 16.61% (37.86% to 31.57%), 14.77% (30.81% to 26.26%), 15.38% (26.26% to 22.22%) and 14.79% (21.29% to 18.14%) in comparison to the conventional cases for the recognition of angry, sad, lombard and happy speech, respectively. Whereas, the non-linear transformation of angry speech using the polynomial function of order $N = 8$ has resulted in the improved speech recognition performances by the decrement in the WERs by 26.81% (37.86% to 27.71%) and 12.23% (31.57% to 27.71%), when compared to the conventional and the linear transformation cases, respectively. For the recognition of sad speech by employing the non-linear transformation using the polynomial function of order $N = 1$, the relative improvement in the recognition performances of 15.84% (30.81% to 25.93%) and 1.26% (26.26% to 25.93%) are achieved in comparison to the performances of conventional and linear transformation-based ASR systems, respectively. Similarly, the non-linear transformation of lombard speech using the polynomial function of order $N = 1$ leads to the deterioration in WERs of 23.72% (26.26% to 20.03%) and 9.85% (22.22% to 20.03%) in comparison to the conventional and the linear transformation cases, respectively. The non-linear transformation of happy speech derived using the polynomial function of order $N = 4$ yields the improvement in the recognition performances by the decrement in WERs of 23.48% (21.29% to 16.29%) and 10.20% (18.14% to 16.29%), when compared to the performances of ASR systems developed in the conventional and the linear transformation cases, respectively. These experimental results illustrate that, the proposed linear transformation of neutral and stressed speech features onto the speech subspace reduces the discrepancy between the different speech units of neutral and stressed speech. Furthermore, the significant improvements in the recognition performances are noted in all the studied stress classes using the proposed non-linear subspace modelling technique in comparison to the stressed speech recognition performances obtained in the conventional and the linear transformation cases. These observations demonstrate that, the polynomial function of specific order effectively exploited the non-linearity between the patterns of neutral and stressed speech and it helps in reducing the variance mismatch between them. Consequently, the non-linear transformation of neutral and stressed speech features onto the speech subspace has resulted in the non-linear features comprising the similar acoustic characteristics. These non-linear features develop the non-linear subspace, whose bases effectively span the speech-specific attributes. The proposed stress normalization method using the non-linear subspace modelling technique through the polynomial function of specific order effectively reduces the non-uniformity between the different speech

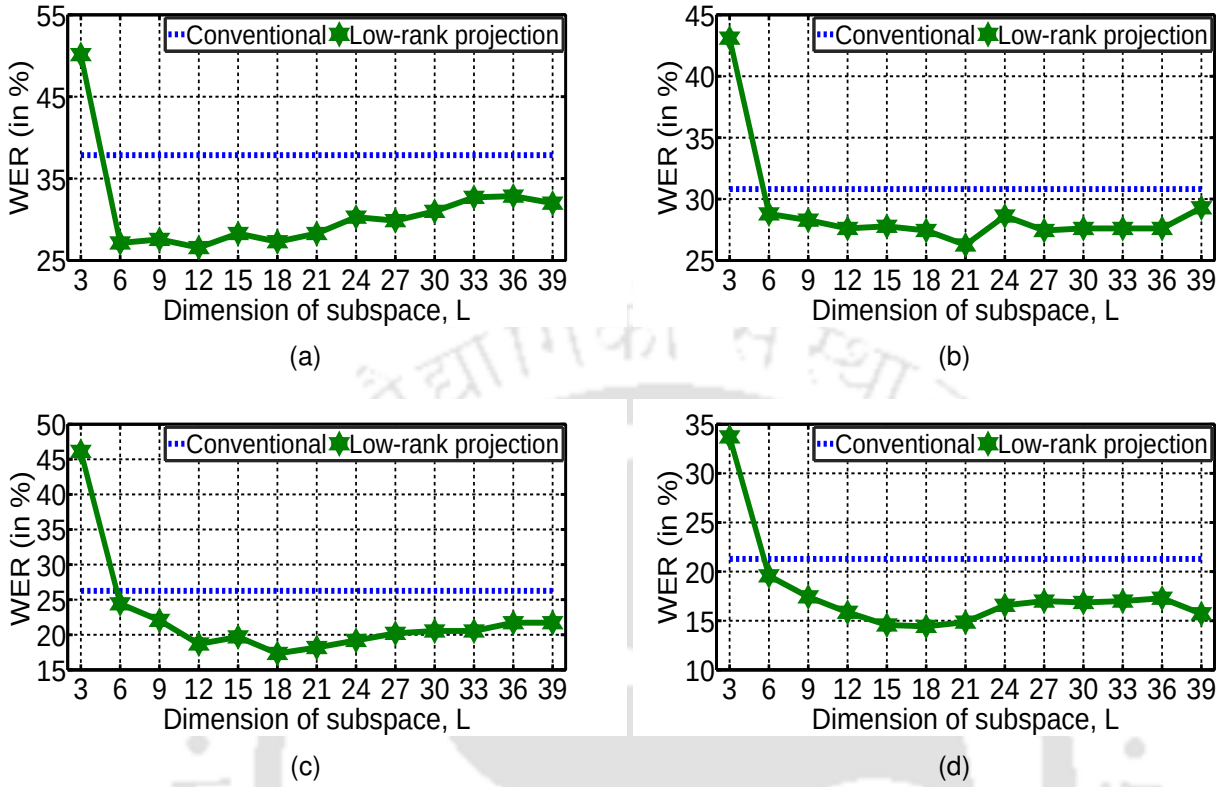


Figure 3.12: Change in the WERs using the proposed stress normalization technique employing the non-linear subspace modelling followed by the HLDA-based low-rank subspace projection method using the TEO-CB-Auto-Env features. (a), (b), (c) and (d) represent the stressed speech recognition performances under angry, sad, lombard and happy conditions, respectively.

units of neutral and stressed speech resulting from stress.

In this work, the effectiveness of the proposed stress normalization method is also investigated onto the subspace having dimension lower than that of the default dimension. The proposed low-rank subspace projection derived using the HLDA-based low-rank subspace projection into the MLLT-based semi-tied adaptation technique is explored only for the best recognition performance cases obtained in each stress classes using the non-linear subspace modelling technique with specific order of polynomial function. The recognition performances for the stressed speech using the non-linear subspace modelling technique utilizing the specific order of polynomial function by varying the dimension of subspace are summarized in the WER-profiles as shown in Figure 3.12. In these figures, WER-profiles are plotted by varying the rank of subspace projection matrix from 3 to 39 in steps of 3. The performances of speech recognition over all the explored stress classes are evaluated for the every reduction in the rank of the projection matrix over the ASR system trained

and tested using the corresponding reduced dimensional decorrelated features of normalized neutral and stressed speech utterances, respectively. The WER values reported in these figures show the significant improvement in the recognition performances, when the feature- and the model-space are adapted into the common subspace constituting the decorrelated characteristics and dimension lower than that of the default dimension. The best recognition performances for the angry, sad, lombard and happy speech are obtained, when the feature- and the model-space are adapted into the decorrelated subspace having specific dimension $L = 12$, $L = 21$, $L = 18$ and $L = 18$ as shown in Figure 3.12(a)$-$Figure 3.12(d), respectively. The maximum relative improvement in the recognition performances using the best performing low-rank features by the decrement in WER values of 29.82% (37.86% to 26.57%), 14.77% (30.81% to 26.26%), 33.97% (26.26% to 17.34%) and 32.22% (21.29% to 14.43%) are obtained under angry, sad, lombard and happy conditions in comparison to the conventional cases, respectively. These improvement in the recognition performances in all the stress classes studied in this work with the deterioration in WERs demonstrates that, the adaptation of features and model parameters into the decorrelated subspace of specific lower dimension leads to less number of parameters for the representation of neutral and stressed speech patterns onto the feature- and model-space, respectively. Consequently, the proposed low-rank subspace projection has resulted in the robust and the precise representation of neutral and stressed speech patterns onto the feature-space as well as the fast convergence for the estimation of model parameters. Moreover, the low-rank subspace projection preserves the speech-specific attributes in the low frequency region and discards the discrepancy induced due to stress in the default-rank features. The acoustic properties of the stressed speech in the low-frequency region become closer to the neutral speech.

Error Analysis: The effectiveness of the proposed stress normalization method employing the non-linear subspace modelling with low-rank subspace projection technique is further measured using the error analysis. As mentioned earlier, the error analysis has been accomplished by exploring the Kullback Leibler (KL) divergence or the relative entropy between the Gaussian subspaces developed using the GMMs comprising the mixture of 32 Gaussian densities [140, 141]. The relative entropy between the Gaussian-subspaces using the proposed method is summarized in the bar plot shown in Figure 3.13. In this plot, the relative entropy values for the conventional case are evaluated by determining the KL divergences between the GMMs trained on the default dimensional features of original neutral and stressed speech utterances. The relative entropies using the non-linear sub-

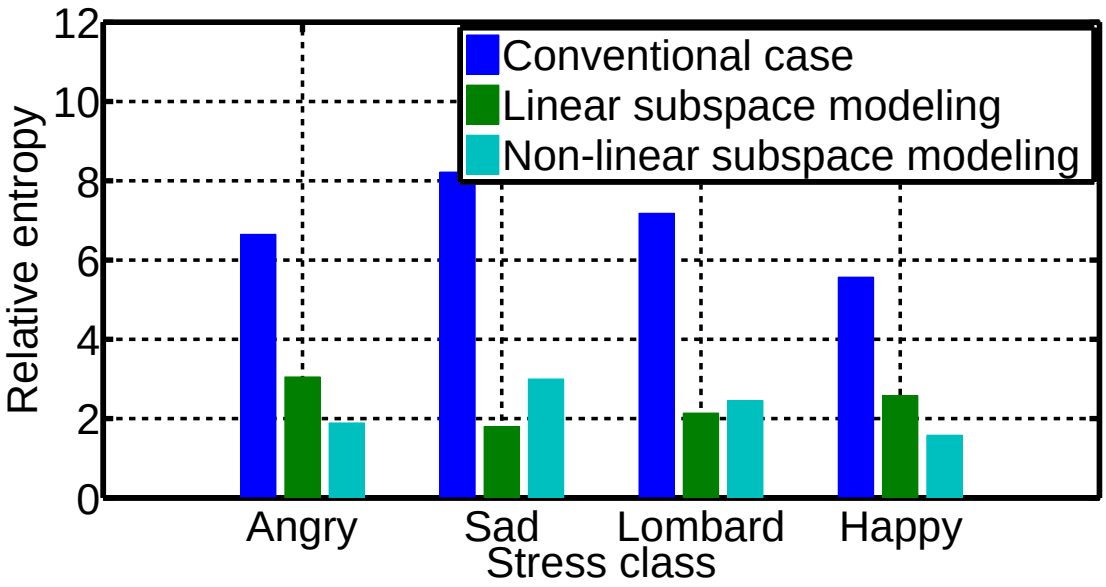
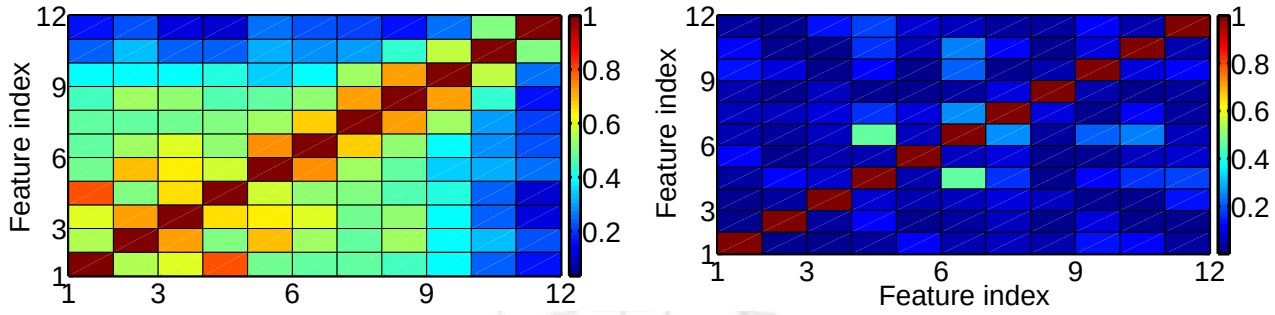


Figure 3.13: The error analysis by exploring the KL divergence metric using the proposed stress normalization technique employing the non-linear subspace modelling followed by the HLDA-based low-rank subspace projection method.

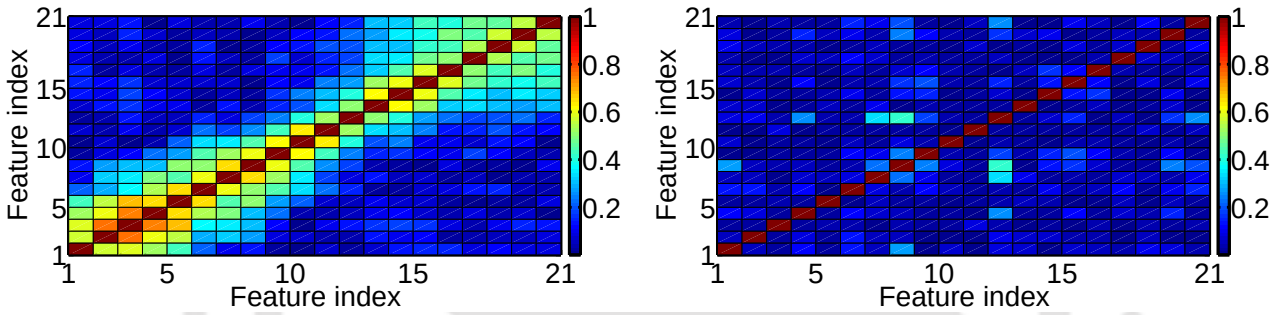
space modelling case are determined by evaluating the KL divergence values between the GMMs trained using the features of neutral and stressed speech, when they are non-linearly transformed onto the speech subspace using the polynomial function of specific order followed by the subspace projection of resulting features onto the specific lower dimensional decorrelated subspace. For comparison purpose, the KL divergence values determined using the proposed orthogonal projection-based linear subspace modelling approach with the low-rank subspace projection onto the specific lower-dimensional decorrelated subspace as given in Figure 3.7 are also summarized in this bar plot. The non-linear transformation of angry, sad, lombard and happy speech using the specific order of polynomial function onto the specific lower dimensional decorrelated subspace lead to the reduction in the KL divergence values by 71.58% (6.65 to 1.89), 63.50% (8.22 to 3), 65.88% (7.18 to 2.45) and 71.63% (5.57 to 1.58) in comparison to the conventional cases, respectively. Moreover, the proposed stress normalization method using the non-linear subspace modelling approach has resulted in the large degree of relative reduction in the KL divergence in comparison to the KL divergence values obtained using the linear subspace modelling technique under angry and happy conditions, respectively. Using non-linear subspace modelling technique, the average value of KL divergences over all the studied stress classes are reduced by 67.68% (6.90 to 2.23) and 6.70% (2.39 to 2.23), when com-

pared to the KL divergences determined using the conventional and the linear subspace modelling derived using the orthogonal projection method, respectively. The degradation in the relative entropy using the proposed method of stress normalization illustrate the reduction in the variance mismatch between the spectral distribution of different speech units of neutral and stressed speech signals. Consequently, Gaussian-subspaces learned using the resulting normalized decorrelated features of neutral and stressed speech constitute the similar acoustic properties.

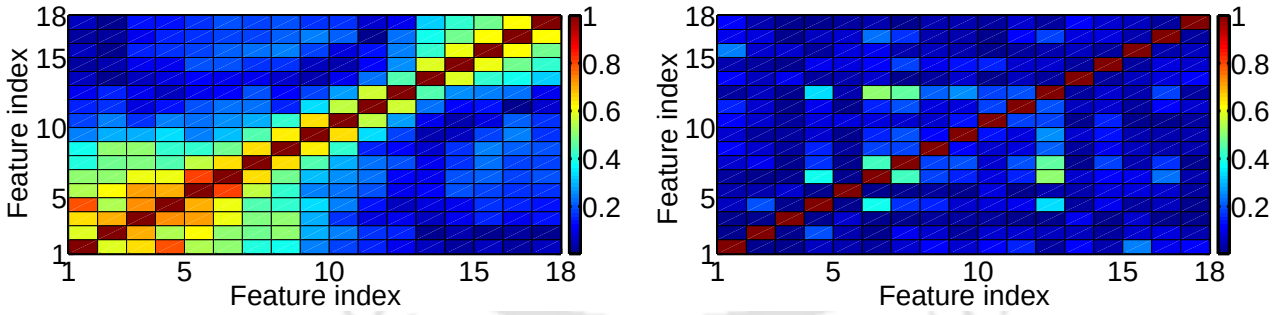
The improved performances of speech recognition and the degraded value of KL divergences for the stressed speech show the effectiveness of the proposed non-linear subspace modelling approach for normalizing the stress-specific attributes. The proposed non-linear transformation of neutral and stressed speech using the polynomial function of specific order is reported to be very effective in reducing the acoustic variabilities between them. In addition to this, the subspace projection of non-linearly transformed features of neutral and stressed speech utterances onto the decorrelated subspace of specific lower dimension leads to the further reduction in the acoustic mismatches between them. To observe the reduction in correlation between the features, the correlation coefficient matrices are plotted as shown in Figure 3.14. In this figure, the correlation coefficient matrices are plotted for the specific lower dimensional decorrelated features of non-linearly transformed features of stressed speech. Additionally, the correlation coefficient matrices are also plotted for the features of specific lower dimension of original stressed speech features derived by retaining only the first specific number of element of feature vectors called as the conventional case as discussed in Subsection 3.3.2. The subspace projection of normalized angry, sad, lombard and happy speech onto the decorrelated subspace of specific lower dimension $L = 12$, $L = 21$, $L = 18$ and $L = 18$ has resulted in the reduction in average values of correlation coefficients of 63.04%(0.46 to 0.17), 62.96%(0.27 to 0.10), 58.06%(0.31 to 0.13) and 62.10%(0.29 to 0.11) in comparison to the average value of correlation coefficients obtained in conventional cases, respectively. These experimental results show that, the proposed HLDA-based low-rank subspace projection significantly reduces the correlation between the features. The lower dimensional decorrelated features of non-linearly transformed neutral and stressed speech create the training and the test environments comprising the similar acoustic characteristics for the ASR system. Furthermore, the adaptation of model-space into the decorrelated subspace having dimension lower than that of the default dimension using the proposed low-rank subspace projection in the MLLT-based semi-tied adaptation technique reduces the parameters



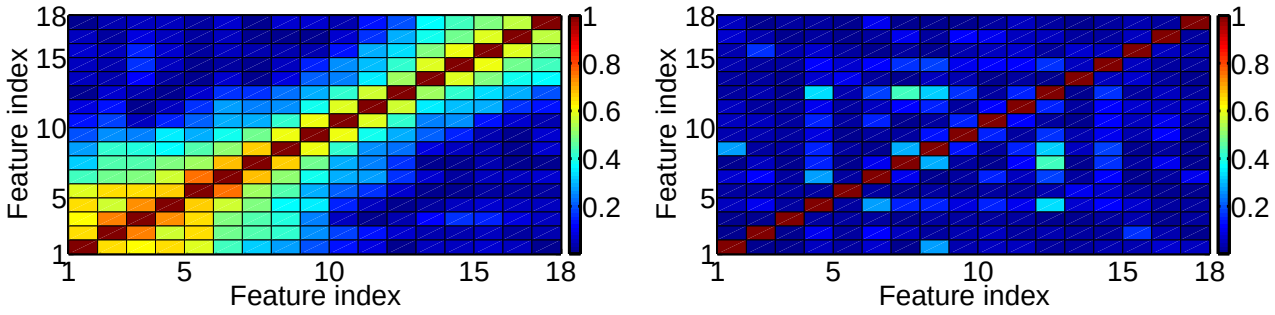
Correlation matrix of $L = 12$ dimensional features; (a) angry speech and (b) normalized angry speech



Correlation matrix of $L = 21$ dimensional features; (c) sad speech and (d) normalized sad speech



Correlation matrix of $L = 18$ dimensional features; (e) lombard speech and (f) normalized lombard speech



Correlation matrix of $L = 18$ dimensional features; (g) happy speech and (h) normalized happy speech

Figure 3.14: The correlation between the specific lower dimensional TEO-CB-Auto-Env features of original stressed speech and normalized stressed speech derived using the proposed non-linear subspace modelling technique along with the HLDA-based low-rank subspace projection method.

for robust modelling. Consequently, ASR systems are appeared very effective for the recognition of stressed speech. The proposed non-linear subspace modelling by exploiting the non-linear transformation using the polynomial function along with the HLDA-based low-rank subspace projection techniques have significantly diminished the variance mismatch between the different speech units of neutral and stressed speech and developed the robust stress normalization algorithm.

3.6 Summary

In this chapter, a novel subspace modelling approach is studied for the development of effective stress normalization algorithms. The linear and the non-linear characteristics between the speech- and the stress-specific attributes have been investigated in the linear and the non-linear subspaces, respectively. The linear subspace is modeled using the orthogonal projection technique by exploiting the linear relationship between the speech and the stress information. Experimental evaluations presented in this work reveal that, the hypothesis of the linearity between the speech- and the stress-specific attributes has resulted in the filtering of an effective speech subspace. The orthogonal projection of stressed speech features onto the specific number of basis vectors of effective speech subspace has been reported very compelling for normalizing the stress information with the significant degree of decrement in the WER values in comparison to the conventional cases. In order to study the non-linear characteristics of stressed speech, the non-linear subspace is modeled by exploiting the non-linear transformation technique using the polynomial function. The effectiveness of non-linear transformation has been investigated by varying the order of polynomials. The experimental observations illustrate that, the subspace projection of neutral and stressed speech onto the speech subspace using the polynomial function of specific order diminishes the variance mismatch between them more pronouncedly in comparison to the subspace projection derived using the linear transformation techniques over the same transformation matrix. Furthermore, the effectiveness of proposed stress normalization algorithms is enhanced by the adaptation of feature- and model-space onto the decorrelated subspace having dimension lower than that of the default dimension. The experimental results show that, the semi-tied adaptation in MLLT framework using HLDA based low-rank subspace projection significantly reduces the correlation between the features and the model parameters. The resulting adapted ASR systems are noted to be very robust for the recognition of stressed speech in all the studied stress classes. After studying the relative performances, it is observed that, the

proposed non-linear subspace modelling technique using the polynomial function of specific order effectively reduces the discrepancy of the spectral distribution of different speech units between the neutral and the stressed speech in comparison to the subspace projection derived using the orthogonal projection and the linear transformation approaches. In the low frequency region, the stressed speech exhibits the characteristics similar to as the neutral speech signal.

4

Stress Normalization Using LPC-Based Posteriorgram Representation With Speaker Adaptation

Contents

The speaker modifies the speech production system to emphasize or de-emphasize the effect of the stressful condition as well as to retain the intelligibility of speech signals [4–8]. The vocal-tract system and the excitation source exhibits different characteristics under stress condition compared to the neutral condition. In Chapter 2, Figure 2.10 and Figure 2.11 have summarized the modified values of frequencies for the formants and the pitch parameters for the stressed speech. The changes in the formants and the pitch frequencies signify the modification in the properties of vocal-tract system and excitation source, respectively. The modified characteristics of vocal-tract system and excitation source alter the spectral distribution of different speech units of stressed speech as shown in Figure 2.3 and Figure 2.4. In literature, the alterations in the parameters of vocal-tract system are investigated in various studies and are reported to be very effective for the stressed speech processing [25, 43–45]. In the work reported in [44, 45], authors have studied the non-linear trend of air propagation within the vocal-tract system under adverse stressful condition. They have introduced the Teager energy operator (TEO)-based processing for the robust representation of non-linear airflow pattern. The proposed non-linear modelling approach has been noted to be very effective for classifying the stress classes. In other work [25], the airflow pattern in the physiological system model was interpreted using the physical parameters of two-mass model and was found to be robust for the classification of stressed speech. Recently, the vocal source Hurst-vectors or pH feature and the binary acoustic mask were observed very effective for emotion classification [43]. These studies demonstrate the significance of vocal-tract system information for the effective processing of stressed speech and motivate us for the investigation on the changes in the vocal-tract system under stress condition for normalizing the stress information. The analysis on the changes in vocal-tract system by exploring the subspace projection-based approaches can provide informative direction for the development of robust and computationally efficient stress normalization technique.

The experimental studies presented in the Chapter 3 have explored the novel subspace projection-based approaches for the reduction of acoustic mismatch between the neutral and the stressed speech signals. In those experiments, the subspace projections were accomplished over the projection matrices learned using the neutral speech data by exploiting the orthogonal projection, the linear transformation and the non-linear transformation techniques to retrieve the acoustic properties of stressed speech similar to the neutral speech. On account of these observations, it is hypothesized that, the projection of vocal-tract system parameters for the stressed speech onto the subspace

consisting of the acoustic parameters of vocal-tract system under neutral condition will significantly alleviate the impact of stress as well as retain the acoustic information. Motivated by these studies, in this Chapter, the aforementioned mismatch between the neutral and the stressed speech is reduced by the subspace projection of acoustic parameters of vocal-tract system for the neutral and the stressed speech onto a common subspace using the posterior probability information. This common subspace has Gaussian features and it consists of vocal-tract system characteristics of neutral speech. The projection of vocal-tract system attributes onto the Gaussian-subspace using the posterior probability information yields the posteriorgram features. In the past few decades, various methodologies in the research area of spoken term detection, speech recognition, language recognition, speaker recognition, speaker identification etc. were effectively developed by studying the posteriorgrams representation of the speech signals [154–160]. These research works illustrate that, the posteriorgram features comprise the significant characteristics of the speech signal. The synthesis of neutral and stressed speech using the corresponding estimated posteriorgram features of vocal-tract system parameters can help in diminishing the acoustic variation between them.

The effectiveness of the proposed stress normalization technique is further improved by suppressing the speaker-specific attributes from both the feature- and the model-space, respectively. Speech signal comprises mainly of phonetic and speaker information. As discussed earlier, speaker modifies the speech production system to acknowledge about the stress environmental factors and to preserve the acoustic information of speech. Both the phonetic- and the speaker-specific attributes of the speech signal are affected under stress condition. The motivation behind the work presented in this chapter is also to search for a subspace that consists of the phonetic information along with the suppressed speaker variabilities. Such a subspace may, in turn, help in reducing the acoustic mismatch between the neutral and the stressed speech. To normalize the speaker variability, we have employed the feature-space maximum-likelihood linear regression (fMLLR) [144, 145, 161] technique in the speaker adaptive training (SAT) [95, 97, 146, 162] framework. Furthermore, reducing the rank of the projection matrix will, in turn, reduce the mismatch in the variances resulting from the stress. Such a low-rank projection matrix can be derived using either the principal component analysis (PCA) [163] or the linear discriminant analysis (LDA) [147] performed on the synthesized neutral speech. It is quite well know that, the LDA-based projections are more effective than those derived using the PCA. This is due to the fact that the former is learned in a supervised manner using

the class labels of the training data. Therefore, LDA based low-rank projections have been explored in this work. The effectiveness of the proposed stress normalization method has been measured using the acoustic modelling techniques based on Gaussian mixture model (GMM) [1], subspace Gaussian mixture model (SGMM) [52–54] as well as deep neutral network (DNN) [55, 56]. The performances are evaluated by exploring the waveform, the spectral distribution, the Kullback Leibler (KL) divergence [140, 141] and the speech recognition for the stressed speech using the two types of databases namely: the Speech under Simulated Stress Condition (SUSSC) database [118] and the Database of German Emotional Speech (Emo-DB) [119], respectively.

The remainder of this chapter is organized as follows: The proposed stress normalization approach using the speech synthesis by exploiting the posteriorgram representation of vocal-tract system parameters is presented in Section 4.1. In Section 4.2, the performance evaluations of the synthesized speech are summarized. Section 4.3 describes the LDA-based low-rank subspace projection technique. The technique for the normalization of speaker variability is presented in Section 4.4. In Section 4.5, the experimental evaluation of the proposed stress normalization method is discussed. Finally, this chapter is summarized in Section 4.6.

4.1 Posteriorgram Representation of Vocal-Tract System Parameters for Speech Synthesis

Stress has salient impact on the vocal-tract system and the excitation source. Consequently, the speech produced under stress condition carries the different acoustic properties in comparison to the speech produced under normal or neutral condition. The acoustic mismatch between the neutral and the stressed speech is reduced by studying the characteristics of vocal-tract system. The changes in the vocal-tract system are investigated by exploring the subspace projection approach. The proposed stress normalization method is summarized in the block diagram shown in Figure 4.1. As depicted in this block diagram, the vocal-tract system parameters for both the neutral and the stressed speech are projected onto a common Gaussian-subspace, which is developed using the vocal-tract system parameters of neutral speech data to reduce the acoustic mismatch between them. In this work, we have accomplished the subspace projection by exploiting the posterior probability information and it has resulted in the posteriorgram representation. The effectiveness of the proposed approach for normalizing the stress-specific attributes depends on the robust estimation of vocal-tract system pa-

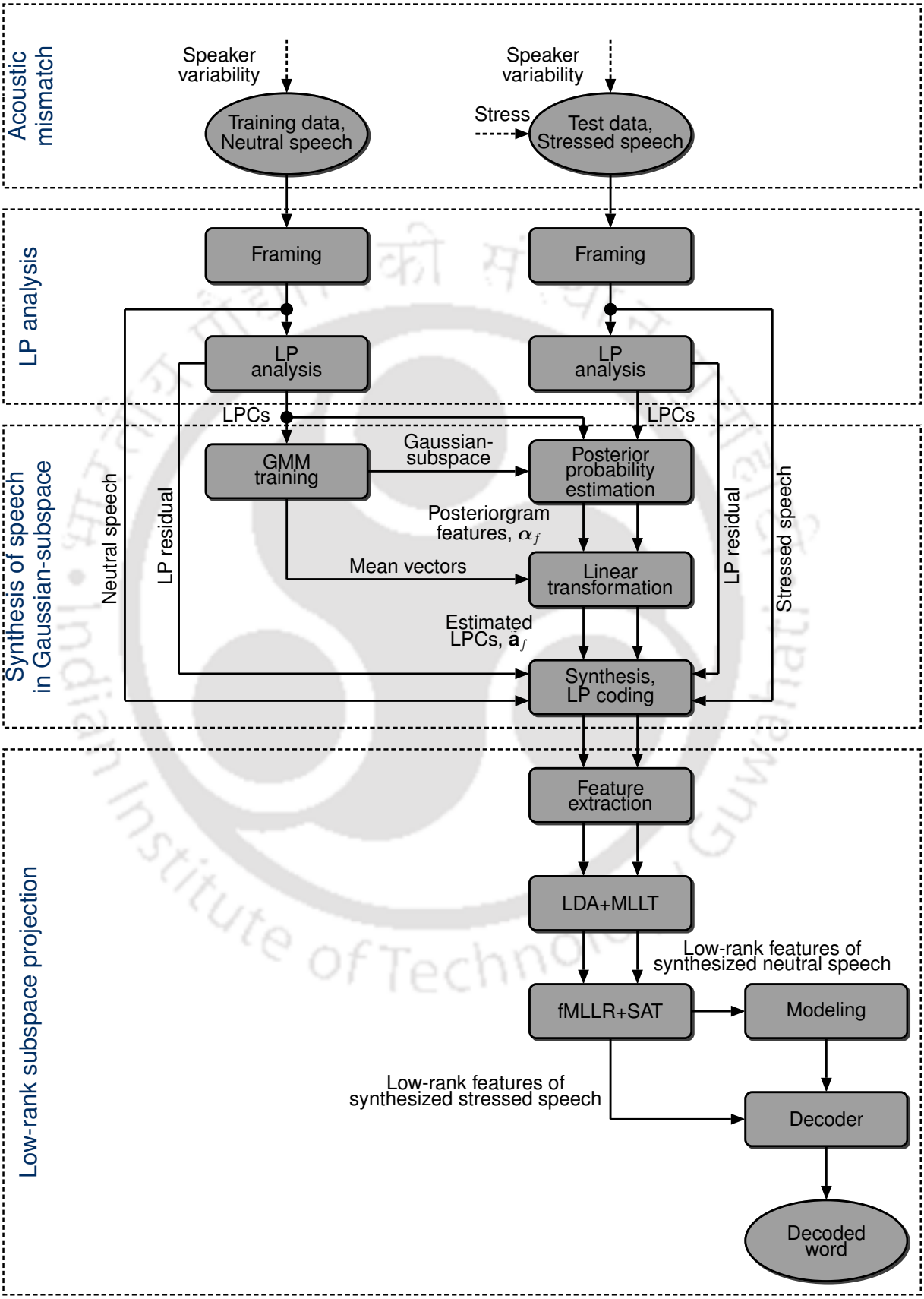


Figure 4.1: Stressed speech recognition using the proposed subspace projection of vocal-tract system parameters onto the Gaussian-subspace followed by the low-rank subspace projection.

rameters for the neutral and the stressed speech. It is hypothesized that, the proposed subspace projection can result in the vocal-tract system parameters with the characteristics similar to the vocal-tract system parameters for the neutral speech. The synthesis of neutral and stressed speech using their corresponding estimated vocal-tract system parameters can help in reducing the variance mismatch between them. In the following Subsection, we have described the methodology of speech synthesis for stress normalization. This is followed by a details on the estimation of robust vocal-tract system parameters using the posteriorgram probability information.

4.1.1 Speech Synthesis

The speech signal is processed through the linear prediction (LP)-based standard source-filter model [1, 101–105,127]. The LP analysis of speech signal extracts the linear prediction coefficients (LPCs) and the residual error, which model the significant characteristics of the vocal-tract system and the excitation source, respectively. These LPCs are influenced by the stress, which leads to the diversified features for the stressed speech compared to the neutral speech features. In this work, both the neutral and the stressed speech are synthesized using their respective estimated LPCs, which constitute the similar acoustic properties as shown in Figure 4.1. The synthesized speech is the speech having normalized stress information and is determined using the LP coding [1, 105, 127] as follows,

$$\tilde{s}_f(n) = e_f(n) - \sum_{p=1}^{P} \tilde{a}_{f_p} s_f(n-p) \tag{4.1}$$

$$e_f(n) = s_f(n) + \sum_{p=1}^{P} a_{f_p} s_f(n-p) \tag{4.2}$$

where, the synthesized speech sample corresponding to the speech sample $s_f(n)$, which belongs to the frame index f of neutral or stressed speech utterance is represented by $\tilde{s}_f(n)$. The stress normalize LPCs are denoted by $\{\tilde{a}_{f_p}\}_{p=1}^{P}$. $\{a_{f_p}\}_{p=1}^{P}$ and $e_f(n)$ represent the LPCs and the residual error of the speech sample $s_f(n)$, determined using the P-order LP analysis. It is evident from Eq. (4.1) and Eq. (4.2) that, the efficacy of proposed approach for the normalization of stress-specific attributes depends on an effective estimation of $\{\tilde{a}_{f_p}\}_{p=1}^{P}$. The following subsection describes the estimation of effective LPCs by exploring their posteriorgram representation.

4.1.2 Estimation of Vocal-Tract System Parameters

The acoustic mismatch between the vocal-tract system for the neutral and the stressed speech are reduced by exploiting the subspace projection technique. To retrieve the acoustic similarities, LPCs for the neutral and the stressed speech utterances are projected onto a common subspace. This common subspace comprises the Gaussian features and is generally referred to as the Gaussian-subspace. The estimation of bases for this common Gaussian-subspace using the LPCs of neutral speech data can help in reducing the aforementioned acoustic mismatch, since they consist of acoustically rich phonetic information. The following steps summarize the proposed subspace projection-based method for the effective estimation of LPCs.

Step I: At first, the set of LPCs, $\mathbf{a}_f = \begin{bmatrix} a_{f_1} & a_{f_2} & \cdots & a_{f_P} \end{bmatrix}^T$ corresponding to each frame of neutral speech utterances belonging to the training database are determined using the P-order LP analysis. In this work, the set of LPCs associated with each frame is referred to as the LP-vector having dimension $P \times 1$. The subscript f represents the frame index. T denotes the transpose operation. In this way, the set of LP-vectors $\{\mathbf{a}_f\}_{f=1}^{N}$ corresponding to all the frames of neutral speech utterances present in the training database are determined. Where N denotes the total frames of neutral speech utterances of training database.

Step II: In the next step, a Gaussian mixture model (GMM) [112] comprising the weighted sum of G-mixture Gaussian densities is trained using the LP-vectors $\{\mathbf{a}_f\}_{f=1}^{N}$ of the neutral speech utterances. The subspace, corresponding to this GMM is considered as the G-dimensional Gaussian-subspace. The mean vectors $\{\mathbf{m}_g^{\alpha}\}_{g=1}^{G}$ corresponding to each Gaussian density capture the vocal-tract system characteristics of the neutral speech.

Step III: For any given LP-vector, $\mathbf{a}_f$ from the set of LP-vectors, $\{\mathbf{a}_f\}_{f=1}^{F}$ of observed neutral or stressed speech utterance having F frames, which belongs to the P-dimensional subspace is projected onto the G-dimensional Gaussian-subspace to $\boldsymbol{\alpha}_f$ as shown in Figure 4.1. Using the information of the projected LP-vector $\boldsymbol{\alpha}_f$, the LP-vector $\tilde{\mathbf{a}}_f$ is estimated by the linear combination of the mean vectors $\{\mathbf{m}_g^{\alpha}\}_{g=1}^{G}$ of the Gaussian densities and is determined as follows,

$$\tilde{\mathbf{a}}_f = \alpha_{f_1}\mathbf{m}_1^{\alpha} + \alpha_{f_2}\mathbf{m}_2^{\alpha} + \cdots + \alpha_{f_G}\mathbf{m}_G^{\alpha} \tag{4.3}$$

where, the coefficients of linear combination are taken as the elements $\{\alpha_{f_g}\}_{g=1}^{G}$ of the pro-

jected LP-vector $\boldsymbol{\alpha}_f$. These elements measure the degree of similarity between the observed LP vector and the Gaussian-subspace and it can help in estimating the LP vector for the neutral and the stressed speech utterances comprising the similar acoustic properties .

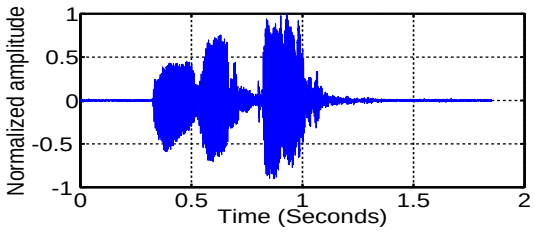
Step IV: The coefficient of linear combination, α_{f_g} with respect to the mean vector $\mathbf{m}_g^{\alpha}$ is determined by measuring the posterior probability P_g, i.e., $\alpha_{f_g} = P_g$ corresponding to the g^{th} Gaussian density for a given $\mathbf{a}_f$ as depicted in Figure 4.1. The summation of posterior probabilities, $\{\alpha_{f_g}\}_{g=1}^{G}$ over all the Gaussian densities is unity i.e., $\sum_{g=1}^{G} \alpha_{f_g} = 1$. In this way, each elements of the projected LP-vector $\boldsymbol{\alpha}_f$ consists of similarity score between the observed LP-vector $\mathbf{a}_f$ and the Gaussian-subspace by measuring the posterior probability corresponding to each Gaussian densities. These posterior probability information is generally referred to as the posteriorgram representation and results in the posteriorgram features [154–160]. These posteriorgram features is used in Eq. (4.3) for the estimation of the effective LP-vector.

The neutral/ stressed speech utterance is synthesized using the elements of the corresponding estimated LP-vectors $\{\tilde{\mathbf{a}}_f\}_{f=1}^{F}$ in Eq. (4.1) and Eq. (4.2). The resulting synthesized neutral and synthesized stressed speech are considered as the speech consisting of similar acoustic properties.

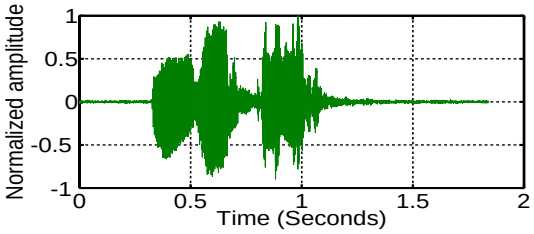
4.2 Evaluation of Synthesized Speech

The proposed stress normalization method has incorporated the synthesis of neutral and stressed speech by exploiting the posteriorgram representation for their corresponding vocal-tract system parameters, respectively. The synthesized neutral and the synthesized stressed speech are considered as the speech having similar acoustic characteristics. In this Section, the quality of synthesized speech signals have been measured onto the two distinct but interconnected frameworks namely: the visual analysis and the error analysis, respectively. In the visual analysis, the synthesized speech signals are quantified by studying the time and the frequency domain representations. The error analysis measures the relative entropy between the Gaussian-subspaces for the evaluation of quality of synthesized speech [140, 141]. In the following paragraphs, we have described the evaluation of synthesized speech signals onto the aforementioned frameworks using the two stressed speech databases, the SUSSC database [118] and the Emo-DB database [119], respectively.

Visual Analysis: The visual analysis investigates the changes in the synthesized speech by interpreting the time- and the frequency-specific characteristics. In time domain, the synthesized speech is evaluated by studying the variations in the air pressure that human auditory system are able to perceive as sound over the time. The quantification of synthesized speech in the frequency domain has been accomplished by analyzing the modification in the distribution of spectral contents with respect to the time parameter. In this experiment, the waveform and the spectrogram are used to represent the variation in the air pressure and the spectral distribution, respectively. Using SUSSC database, the waveforms and the spectrograms are plotted for the synthesized speech utterances of word /angoothi/ recorded under neutral, angry, sad, lombard and happy conditions using the same speaker as shown in Figure 4.2 and Figure 4.3, respectively. Figure 4.4 and Figure 4.5 have summarized the waveforms and the spectrograms for the synthesized speech utterances of sentence /Tonight I could tell him/ of Emo-DB database recorded from the same speaker under neutral, angry, sad, happy and boredom conditions, respectively. For comparison purpose, the waveforms and the spectrograms are also plotted for the original raw speech utterances in all the studied cases. It has been noted that, the waveforms of synthesized speech utterances are quite similar to as the waveforms of original raw speech utterances with the slight variation in the amplitude values of air pressures. This behavior is consistently exhibited for the neutral and the stressed speech utterances of both the databases explored in this work as shown in Figure 4.2 and Figure 4.4, respectively. As depicted from Figure 4.2(c), Figure 4.2(d), Figure 4.4(c) and Figure 4.4(d), the changes in the air pressure are observed approximately between 0.75 to 1 and 2 to 2.5 seconds in the original raw angry speech and the synthesized angry speech of SUSSC and Emo-DB databases, respectively. Similarly, the proposed synthesis process reduces the amplitude values of air pressure of sad speech of both the explored databases as shown in Figure 4.2(e), Figure 4.2(f), Figure 4.4(e) and Figure 4.4(f), respectively. These quite similar variations in the sound pressure for the original raw speech and the synthesized speech utterances illustrate that, the proposed synthesis process preserve the original phonetic structures and it helps in retaining the intelligibility of speech signals. Whereas, the little differences in the sound pressure demonstrate that, the posteriorgram representation of vocal-tract system parameters for the neutral and the stressed speech reduces the variance mismatch between them. Consequently, the synthesis of neutral and stressed speech using their corresponding estimated vocal-tract system parameters helps in constituting the similar acoustic properties while maintaining their original patterns over the

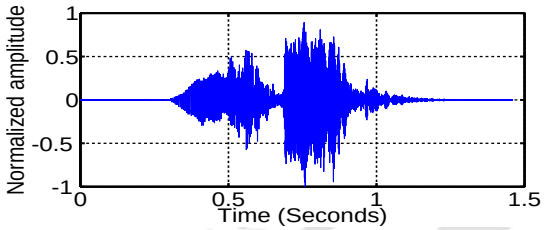


(a)

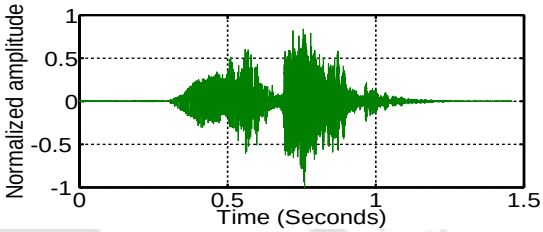


(b)

Speech recorded under neutral condition: (a) original speech and (b) synthesized speech

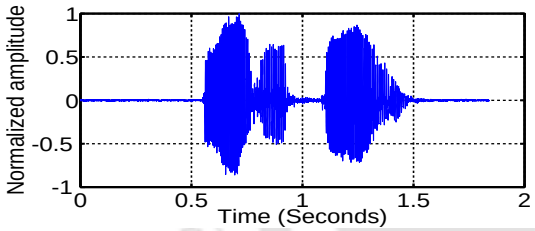


(c)

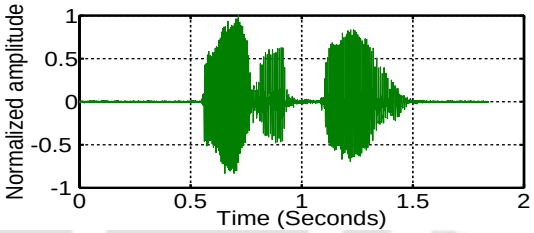


(d)

Speech recorded under angry condition: (c) original speech and (d) synthesized speech

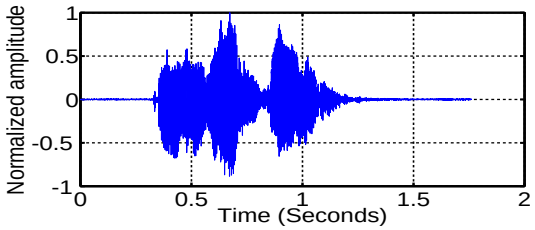


(e)

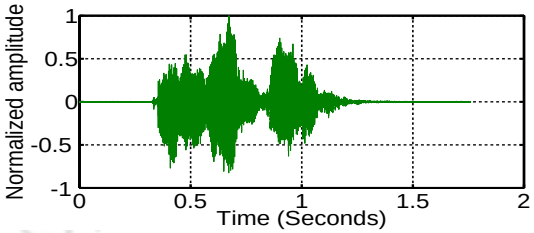


(f)

Speech recorded under sad condition: (e) original speech and (f) synthesized speech

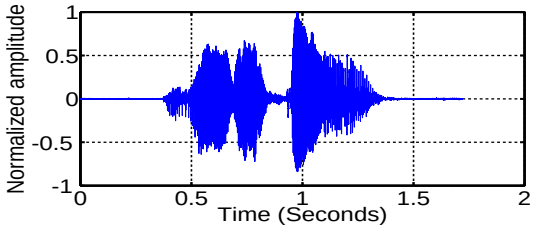


(g)

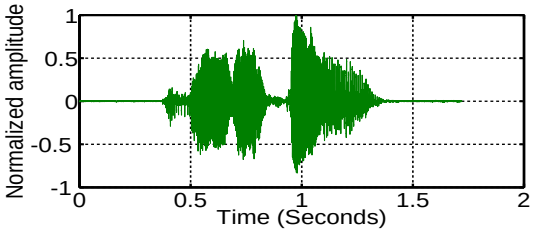


(h)

Speech recorded under lombard condition: (g) original speech and (h) synthesized speech



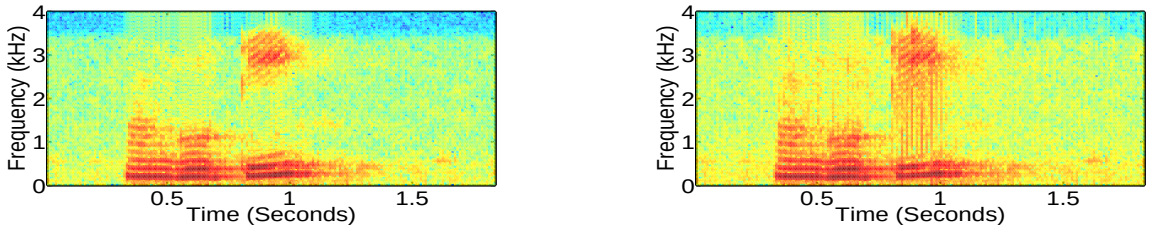
(i)



(j)

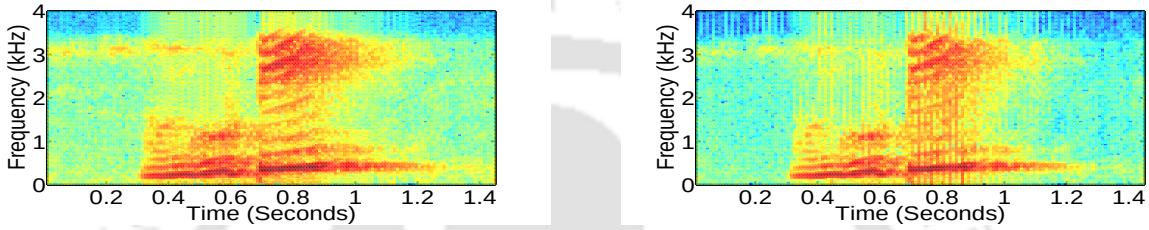
Speech recorded under happy condition: (i) original speech and (j) synthesized speech

Figure 4.2: The visual analysis of speech utterances of word /angoothi/ of SUSSC database by plotting their waveforms. The waveforms are plotted for the original raw speech and the synthesized speech determined using the proposed stress normalization method.



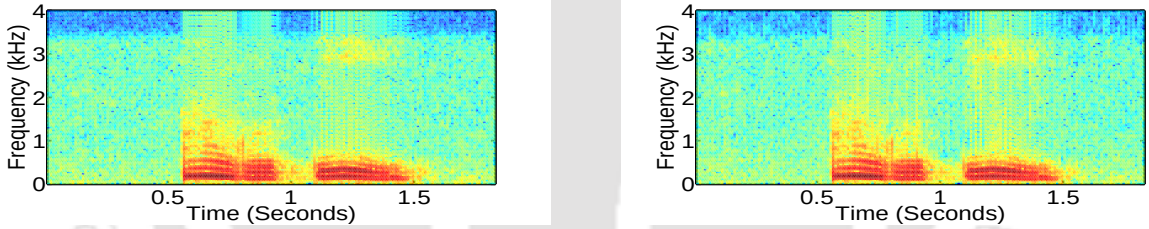
(a) (b)

Speech recorded under neutral condition: (a) original speech and (b) synthesized speech



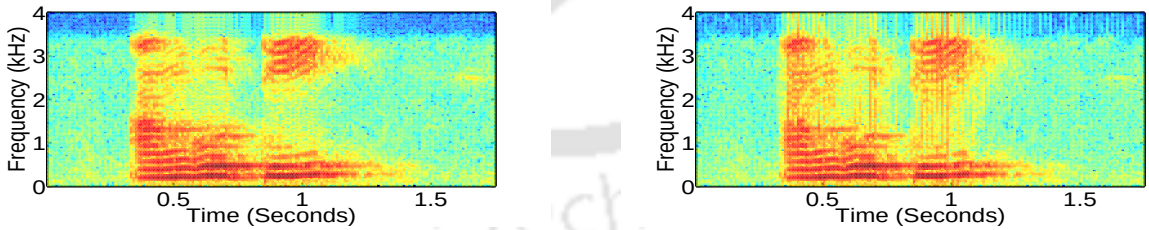
(c) (d)

Speech recorded under angry condition: (c) original speech and (d) synthesized speech



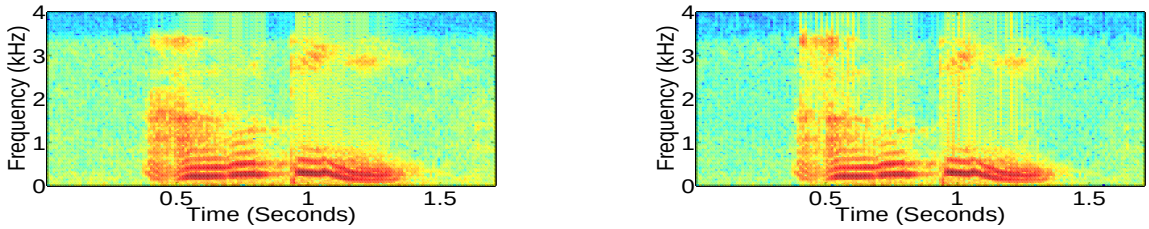
(e) (f)

Speech recorded under sad condition: (e) original speech and (f) synthesized speech



(g) (h)

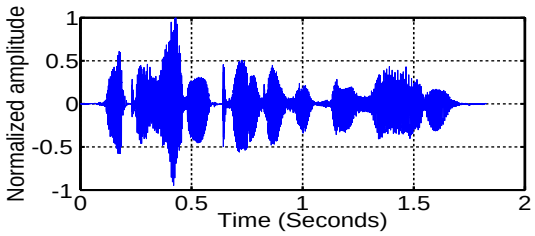
Speech recorded under lombard condition: (g) original speech and (h) synthesized speech



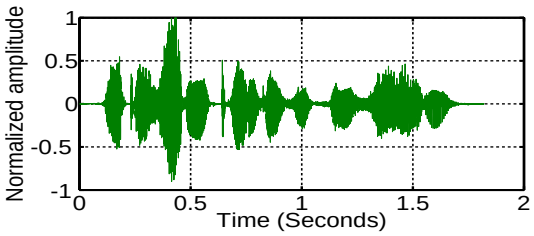
(i) (j)

Speech recorded under happy condition: (i) original speech and (j) synthesized speech

Figure 4.3: The visual analysis of speech utterances of word /angoothi/ of SUSSC database by plotting their spectrograms. The spectrograms are plotted for the original raw speech and the synthesized speech determined using the proposed stress normalization method.

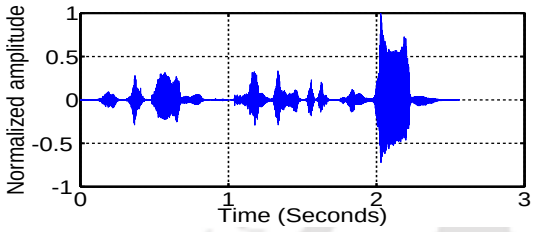


(a)

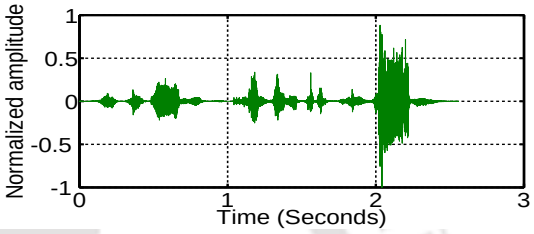


(b)

Speech recorded under neutral condition: (a) original speech and (b) synthesized speech

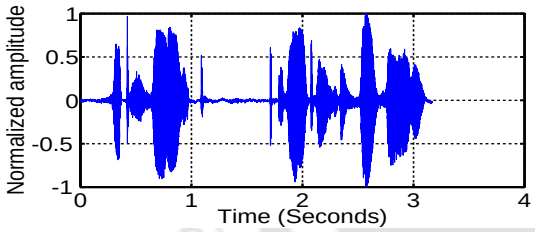


(c)

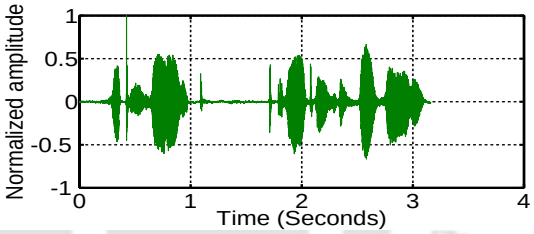


(d)

Speech recorded under angry condition: (c) original speech and (d) synthesized speech

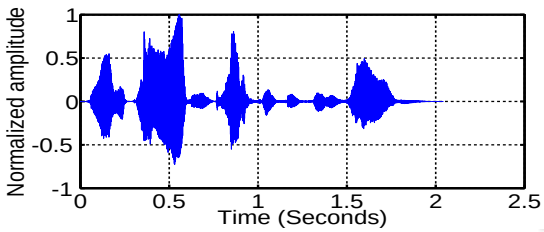


(e)

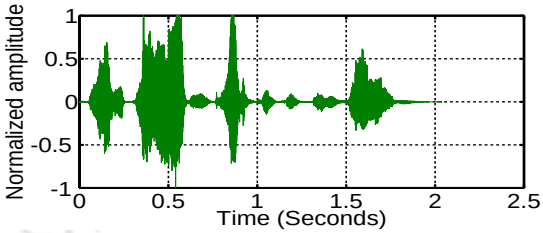


(f)

Speech recorded under sad condition: (e) original speech and (f) synthesized speech

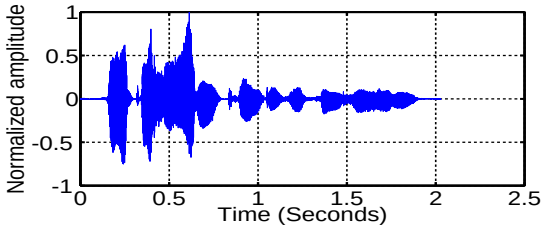


(g)

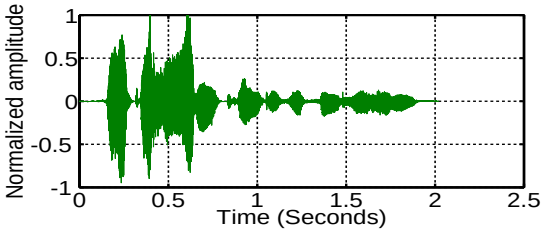


(h)

Speech recorded under happy condition: (g) original speech and (h) synthesized speech



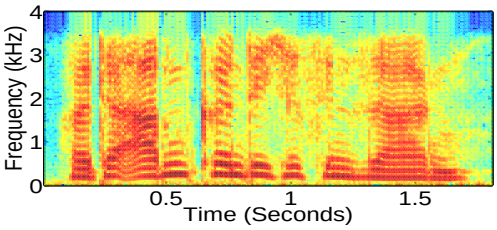
(i)



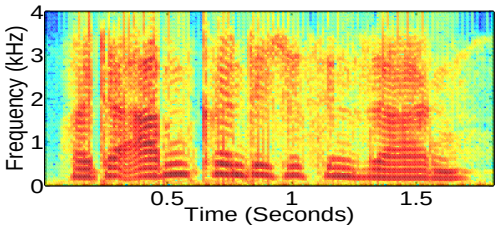
(j)

Speech recorded under boredom condition: (i) original speech and (j) synthesized speech

Figure 4.4: The visual analysis of speech utterances of sentence /Tonight I could tell him/ of Emo-DB database by plotting their waveforms. The waveforms are plotted for the original raw speech and the synthesized speech determined using the proposed stress normalization method.

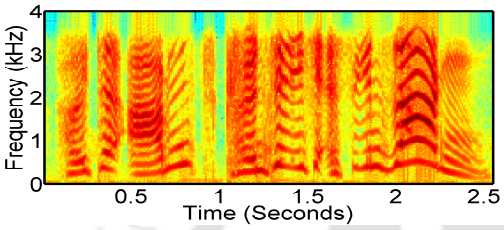


(a)

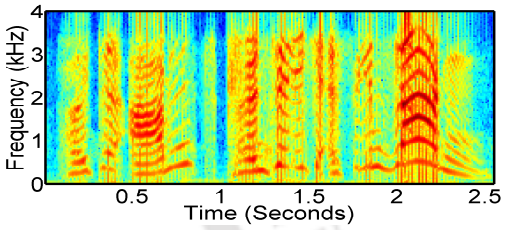


(b)

Speech recorded under neutral condition: (a) original speech and (b) synthesized speech

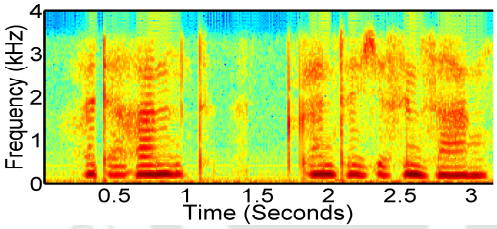


(c)

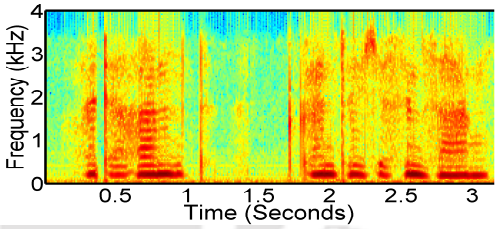


(d)

Speech recorded under angry condition: (c) original speech and (d) synthesized speech

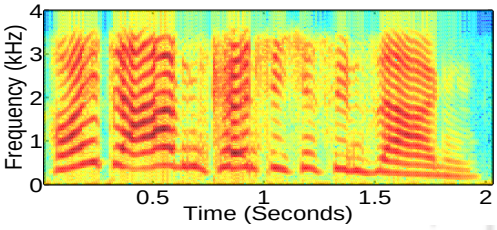


(e)

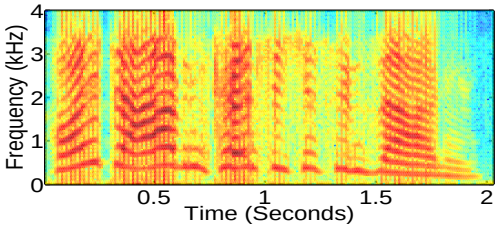


(f)

Speech recorded under sad condition: (e) original speech and (f) synthesized speech

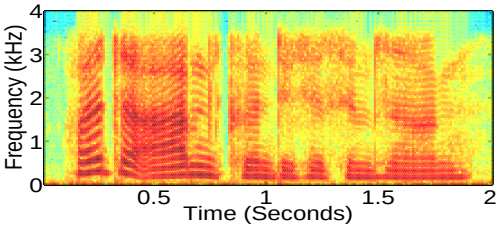


(g)

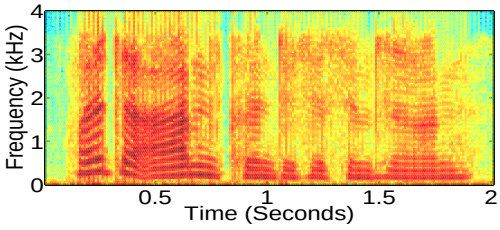


(h)

Speech recorded under happy condition: (g) original speech and (h) synthesized speech



(i)



(j)

Speech recorded under boredom condition: (i) original speech and (j) synthesized speech

Figure 4.5: The visual analysis of speech utterances of sentence /Tonight I could tell him/ of Emo-DB database by plotting their spectrograms. The spectrograms are plotted for the original raw speech and the synthesized speech determined using the proposed stress normalization method.

time. In frequency domain, the larger modifications in the spectral distribution of synthesized speech are found in the moderate to high audio frequency bandwidth in comparison to the spectral contents of original raw speech. The similar changes in the spectral properties of stressed speech were also observed in the studies reported in [6]. The maximal changes in the spectral distribution have been observed between 1 kHz to 3.5 kHz audio frequency bandwidth for the synthesized lombard speech, when compared to the original raw lombard speech as shown in Figure 4.3(g) and Figure 4.3(h). Whereas, the larger modifications are appeared between 1.5 kHz to 3.5 kHz audio frequency range for the synthesized sad speech in comparison to the original raw sad speech over both the explored databases as depicted in Figure 4.3(e), Figure 4.3(f), Figure 4.5(e) and Figure 4.5(f), respectively. In case of happy speech, using both set of databases, the proposed synthesis method changes the spectral properties between 1.5 kHz to 3.5 kHz audio frequency band as shown in Figure 4.3(i), Figure 4.3(j), Figure 4.5(g) and Figure 4.5(h), respectively.These alteration in the spectral distribution of the synthesized stressed speech utterances revealed that, the proposed synthesis method helps in suppressing the stress-specific attributes generally occur in the moderate to high frequency regions and also helps in preserving the original phonetic characteristics.

Error Analysis: In the error analysis, the synthesized speech signals are quantified by measuring the relative entropy between the Gaussian-subspaces developed using the synthesized neutral and the synthesized stressed speech utterances, respectively. The relative entropy is measured by exploring the Kullback Leibler (KL) divergence metric. Figure 4.6 summarizes the KL divergence values determined over both the SUSSC and the Emo-DB databases using the bar plots. The Gaussian-subspaces are created using the GMMs consisting of mixture of 32 diagonal Gaussian densities. The GMMs are trained using 2322, 700, 594, 594 and 700 utterances of neutral, angry, sad, lombard and happy speech of SUSSC database, respectively. Using Emo-DB database, the GMMs are developed on 79, 127, 62, 71, 69, 46 and 81 utterances of angry, sad, happy, fear, disgust and boredom speech, respectively. In Figure 4.6(b), the labeling of X-axis 'Disg.' and 'Bore.' represent the disgust and the boredom condition, respectively. For computing speech features, the speech signals are analyzed using a 20 msec Hamming window and frame rate of 10 msec. A 21-channel Mel-filterbank is employed to compute the 13-dimensional base MFCCs ($C0 - C12$) features [100] for each frame of speech. In this figure, the relative entropy values for the conventional case are evaluated by determining the KL divergence between the GMMs trained using the features of original raw neutral and

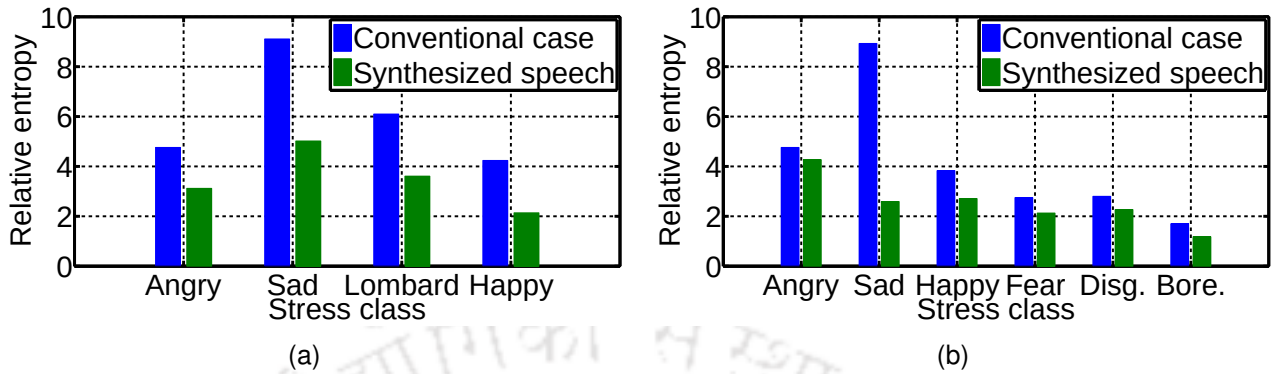


Figure 4.6: Evaluation of synthesized speech determined using the proposed stress normalization by measuring the relative entropy between the Gaussain-subspaces using the KL divergence metric. (a) and (b) represent the KL divergence values with respect to the SUSSC and the Emo-DB databases, respectively.

stressed speech utterances, respectively. The relative entropy values in case of synthesized speech are measured between the Gaussian-subspaces developed on the features of synthesized neutral and synthesized stressed speech, respectively. The synthesis of speech utterances of angry, sad, lombard and happy speech of SUSSC database using the proposed method has resulted in the reduction in the KL divergence values of 34.45% (4.76 to 3.12), 44.89% (9.11 to 5.02), 40.82% (6.10 to 3.61) and 49.53% (4.24 to 2.14), when compared to the KL divergences determined in conventional cases, respectively, as shown in bar plots in Figure 4.6(a). Similarly, using Emo-DB database, when the speech signals are synthesized using the proposed stress normalization method, the KL divergence values for angry, sad, happy, fear, disgust and boredom speech are reduced by 10.30% (4.76 to 4.27), 71% (8.93 to 2.59), 29.24% (3.83 to 2.71), 22.54% (2.75 to 2.13), 18.93% (2.80 to 2.27) and 30.59% (1.70 to 1.18) in comparison to the KL divergences determined in the conventional cases, respectively, as depicted in Figure 4.6(b). Moreover, the proposed stress normalization method has reduced the relative entropies between the Gaussian-subspaces over all the studied stress classes by average value of 42.64% (6.05 to 3.47) and 38.98% (4.13 to 2.52) using the SUSSC and the Emo-DB databases in comparison to the average values of KL divergences obtained in conventional cases, respectively. These reduced values of KL divergences illustrate that, the proposed stress normalization method by exploiting the posteriorgram representation reduces the acoustic mismatch between the vocal-tract system parameters of neutral and stressed speech. Consequently, the Gaussian-subspaces developed using the synthesized neutral and the synthesized stressed speech, which are

determined by incorporating their corresponding estimated vocal-tract system parameters comprises the similar acoustic characteristics with reduced values of relative entropy.

4.3 The LDA-Based Low-Rank Subspace Projection

In this Section, the effectiveness of the proposed stress normalization method is investigate onto the lower dimensional subspace by introducing the low-rank subspace projection approach. In Chapter 3, Figure 3.5−Figure 3.7, Figure 3.12 and Figure 3.13 have manifested that, the adaptation of feature- and model-space onto the decorrelated subspace having dimension lower than that of the default dimension and consisting of decorrelated features of neutral speech data has resulted in the better accuracy for tasks involving the recognition of stressed speech. These observations illustrate that, the proposed low-rank subspace projection effectively reduces the acoustic mismatch between the neutral and the stressed speech. Motivated by these studies, we intended to enhance the effectiveness of the proposed stress normalization method by the adaptation of feature- and model-space onto the another common subspace, whose bases can effectively span the decorrelated features of synthesized neutral speech utterances. In this work, we have explored the linear discriminant analysis (LDA)-based low-rank subspace projection [147] in the maximum likelihood linear transformation (MLLT)-based semi-tied adaptation [148] framework to reduce the dimension as well as to decorrelate the feature- and the model-space as shown in Figure 4.1. The low-rank subspace projection matrix is learned using the synthesized neutral speech utterances. The subspace projection of both the synthesized neutral and the synthesized stressed speech utterances using this projection matrix captures the principal dimensions of the acoustic variations represented by the synthesized neutral speech data and it can help in improving the recognition performances for the stressed speech. In the following, the technique used for the low-rank subspace projection is described.

Consider the K-dimensional synthesized neutral speech feature vector $\mathbf{x}_t$ at time t. These features are time-spliced considering the first t_f frames surrounding the current frames. The resulting C-dimensional ($C = K(2t_f + 1)$) feature vector $\mathbf{x}_{t_C}$ happens to be richer in terms of the context information. The computational complexity of the acoustic modelling developed using the time-spliced features gets enhanced significantly. Therefore, the dimension of features is generally reduced by exploiting the LDA-based low-rank subspace projection technique [147]. The LDA is used to reduce

the dimension of features to L, $(L \leq C)$ as follows,

$$\mathbf{z}_{t_C} = \mathbf{L}\mathbf{x}_{t_C} = \begin{bmatrix} \mathbf{L}_L \mathbf{x}_{t_C} \\ \mathbf{L}_{C-L} \mathbf{x}_{t_C} \end{bmatrix} = \begin{bmatrix} \mathbf{z}_{t_L} \\ \mathbf{z}_{t_{C-L}} \end{bmatrix} \tag{4.4}$$

where, $\mathbf{L}_L$ and $\mathbf{L}_{C-L}$ contain the first L and the remaining $(C-L)$ rows of the projection matrix $\mathbf{L}$ having dimension $C \times C$, respectively. The feature vectors, $\mathbf{z}_{t_L}$ and $\mathbf{z}_{t_{C-L}}$ comprise the first L and the remaining $(C-L)$ elements of linearly transformed features $\mathbf{z}_{t_C}$, respectively. The vector $\mathbf{z}_{t_L}$ forms L dimensional decorrelated feature vector. Using $\mathbf{L}_L$, the synthesized neutral as well as the synthesized stressed speech features are projected onto the decorrelated feature subspace having dimension lower than that of the default dimension. Moreover, the MLLT-based semi-tied adaptation technique is employed to further decorrelate the resulting features and to develop the decorrelated model-space. In general, the final feature dimension L is selected as 39 or 40 for the ASR system. As discussed earlier, the stressed speech exhibits a greater variance in comparison to the neutral speech leading to a degraded performance. Consequently, learning the subspace transformation matrix $\mathbf{L}_L$ on the synthesized neutral speech to reduce the feature dimensions below 40, will help alleviate the variance mismatch. During training, the synthesized neutral speech features are projected onto the lower dimensional subspace before learning the parameters of the acoustic model as depicted in Figure 4.1. For decoding, the synthesized stressed speech features are also projected onto the decorrelated feature subspace using the same projection matrix used in training.

4.4 Normalization of Speaker Variability

The speaker changes the movement of anatomical structure of speech production system to express the condition of stressful environment and to preserve the acoustic information of speech signals. As discussed in Chapter 1, speakers with age and gender variations produce the same speech utterances under the similar stress condition with the varying parameters of stress. In literature, numerous works have been manifested that, the speaker-specific variability appears pronouncedly in the stressed speech [4–8]. The speaker variability causes high overlap between the different speech units of neutral and stressed speech. Consequently, the feature- and the model-space of speaker independent (SI) automatic speech recognition (ASR) system developed using the neutral speech data exhibit high variance for the recognition of stressed speech. Therefore, the normalization of speaker

information from both the neutral and the stressed speech will reduce the variance mismatch resulting from speakers under stress conditions. The adaptation of ASR system using the speaker normalize speech features will significantly diminish the acoustic mismatch between the training and the test environments and it will lead to the robust ASR system against the users under stress condition. The following paragraph summarizes a brief discussion on the method used for normalizing the speaker variability. This is followed by the evaluation of speaker normalization for the stressed speech.

Method of Speaker Normalization: In order to address the speaker variability, we have employed the feature-space maximum-likelihood linear regression (fMLLR)-based speaker normalization technique [144, 145, 161]. The decorrelated features $\mathbf{z}_{t_L}^{\varphi}$ from a particular speaker φ resulting from the LDA+MLLT as discussed in Section 4.3 is adapted to $\hat{\mathbf{z}}_{t_L}^{\varphi}$ through the fMLLR transformation matrix $\mathbf{W}^{\varphi}$ and is determined as follows

$$\hat{\mathbf{z}}_{t_L}^{\varphi} = \mathbf{W}^{\varphi} \boldsymbol{\xi}_{t_L}^{\varphi} \tag{4.5}$$

where, $\boldsymbol{\xi}_{t_L}^{\varphi} = \begin{bmatrix} 1 & \mathbf{z}_{t_L}^{\varphi\ T} \end{bmatrix}^T$ constitutes the feature vector $\mathbf{z}_{t_L}^{\varphi}$ along with the unity element. The employed fMLLR transformation $\mathbf{W}^{\varphi}$ having dimension $(L \times (L+1))$ is learned in the speaker adaptive training (SAT) framework [95, 97, 146, 162] using the decorrelated features of synthesized neutral speech utterances. The speaker adaptive training adapts the speaker independent speech recognition system into the speaker dependent (SD) speech recognition system using low amount of adaptation data. In the works reported in [94–99], the speaker dependent speech recognition system are found to be very effective with improved recognition performance for the unknown test speakers. In addition to this, the features obtained by the concatenation of the discussed transformations (LDA + MLLT + fMLLR) are reported to be very effective in the case of DNN-based ASR systems [99, 164].

Evaluation of Speaker Normalization: The effectiveness of speaker normalization for reducing the acoustic mismatch between the neutral and the stressed speech is measured by evaluating the performances of stressed speech recognition tasks using the SUSSC database [118] as discussed in Section 2.1 in Chapter 2. All the experimental results of this experiment are evaluated using the experimental setup similar to as described in Subsection 2.3.2 in Chapter 2. The recognition performances for the stressed speech after normalizing the speaker variability using the 39-dimensional TEO-CB-Auto-Env and the 39-dimensional- MFCC features with respect to the GMM-HMM, the

Table 4.1: The recognition performances for stressed speech (WER in %) by empoying the speaker normalization method over the MFCC and the TEO-CB-Auto-Env features with respect to GMM-HMM, SGMM-HMM, DNN-HMM and DNN-SGMM systems.

Acoustic modelling approach	Stress class	WER (in %)				% Relative reduction	
		Conventional case		Speaker adaptation			
		Feature type		Feature type		Feature type	
		MFCC	TEO-CB-Auto-Env	MFCC	TEO-CB-Auto-Env	MFCC	TEO-CB-Auto-Env
GMM-HMM	Neutral	0.71	0	0	0	100	NI
	Angry	25.71	41.71	7.43	7.71	71.10	81.51
	Sad	7.41	31.31	3.03	6.40	59.11	79.56
	Lombard	6.73	24.07	1.52	3.20	77.41	86.70
	Happy	7.86	28.29	3.14	6	60.05	78.79
SGMM-HMM	Neutral	0.14	0	0	0	100	NI
	Angry	48.71	44.43	15.71	5	67.75	**88.75**
	Sad	24.07	36.36	10.77	3.70	55.25	**89.82**
	Lombard	26.26	28.79	6.06	1.35	76.92	**95.31**
	Happy	28.14	33	9.86	3.57	64.96	**89.18**
DNN-HMM	Neutral	1.29	0.71	0.29	0.14	77.52	80.28
	Angry	18.86	11.29	7	3.71	62.88	67.14
	Sad	9.09	12.96	1.35	3.37	85.15	74
	Lombard	8.42	7.41	2.53	2.02	69.95	72.74
	Happy	8.57	7.29	2.14	3.29	75.03	54.87
DNN-SGMM	Neutral	1.71	0.86	1.14	0.43	33.33	50
	Angry	24.29	12.57	21.43	12.57	11.77	0
	Sad	16.33	13.13	15.15	9.93	7.23	24.37
	Lombard	16.16	7.74	12.12	4.71	25	39.15
	Happy	13.71	7.57	11.29	6.71	17.65	11.36

SGMM-HMM, the DNN-HMM and the DNN-SGMM systems are summarized in Table 4.1. The DNN-HMM and the DNN-SGMM systems constitute the DNN comprising 3 hidden layers to estimate the posterior probability over their states. The WERs reported in this table for the conventional cases correspond to the case of using baseline ASR system developed on the features of original raw neutral speech data and tested using the original raw neutral and stressed speech utterances. The WER values for the speaker adaptation case are obtained from the 40-dimensional features of the original raw speech signals. These 40-dimensional features are derived from the time splicing of context size of 7 i.e. $\pm$ 3 frames, surrounding to the current frame of the speech. After that, LDA-based low-rank subspace projection is employed in MLLT-based semi-tied adaptation framework to reduce the dimension upto 40 and to decorrelate the feature- and the model-space. This is followed by the fMLLR-based linear regression transformation in SAT framework for the normalization of speaker variability. In this table, the bold face WER values and the acronyms 'NI' denote the best speech recognition performances and no relative improvement in the speech recognition performances in comparison to the conventional cases, respectively. The WERs depicted in this table show that the, speaker normalization significantly improves the recognition performances in comparison to the conventional cases in all the explored stress classes. This improvement is consistent across all the feature types as well as the acoustic modelling approaches. After normalizing the speaker information, the SGMM-HMM system developed on the TEO-CB-Auto-Env features has resulted in the best recognition performances with maximum relative decrement in the WERs of 88.75%, 89.82%, 95.31% and 89.18%, when compared to the WERs obtained in conventional cases under angry, sad, lombard and happy conditions, respectively. These experimental results demonstrate the similar phonetic structure for the training and the testing environments of speech recognition system, when trained and tested using the speaker normalize features of neutral and stressed speech, respectively. Speaker normalization reduces the variance mismatch between the speech units of neutral and stressed speech and helps in creating the similar characteristics for the training and the test environments for the speech recognizer. The resulting ASR system has leaded to the improved the accuracy of speech recognition for recognizing the stressed speech.

These observations motivate us to normalize the speaker variability from the synthesized speech signals. The normalization of speaker information from both the synthesized neutral and the synthesized stressed speech will additionally reduce the variance mismatch resulting from different speak-

ers. The speaker normalization can help in further enhancing the effectiveness of the proposed stress normalization technique. In this work, the speaker variability is normalized from both the feature- and the mode-space using the aforementioned methodology as discussed in above paragraph for increasing the robustness of proposed stress normalization method as shown in Figure 4.1.

4.5 Results and Discussion

In this Section, the proposed stress normalization method employing the LPC-based posteriorgram representation with speaker adaptation technique is evaluated onto the framework of speech recognition for the stressed speech. A word error rate (WER) metric is used to measure the performance of various ASR systems developed in this work and is determined using the Eq. 2.9 as illustrated in Chapter 2. In the following paragraphs, the experimental setup used and the performance evaluation of the proposed stress normalization method are described.

Experimental Setup: The performance of the proposed stress normalization method is evaluated using the SUSSC database reported in [118] as described in Section 2.1 in Chapter 2. In modelling paradigm, the GMM-HMM, the SGMM-HMM, the DNN-HMM and the DNN-SGMM systems are developed on the 39-dimensional Mel-frequency cepstral coefficients (MFCCs) [100] and 39-dimensional TEO-CB-Auto-Env [44, 45] features. The vocal-tract system parameters are extracted using the $P = 25$ order LP analysis [1, 101–105]. The Gaussian-subspace of dimension $G = 64$ is created using the mixture of 64 diagonal Gaussian densities in GMM. The remaining experimental setup is similar to as described in Subsection 2.3.2 in Chapter 2.

Performance Evaluation: The recognition performances of stressed speech by exploring the proposed stress normalization method using the MFCC and the TEO-CB-Auto-Env features over the four types of acoustic modelling paradigms namely: the GMM-HMM, the DNN-HMM, the SGMM-HMM and the DNN-SGMM systems are summarized in Table 4.2−Table 4.5, respectively. In these tables, WERs reported in conventional case are determined using the baseline ASR systems. As discussed in Section 4.4, the baseline ASR systems are trained on the 39-dimensional features of original raw neutral speech utterances. The WER values summarized in case of using the proposed method are obtained from the 40-dimensional features of the synthesized speech signals, which are determined from the concatenation of context size of 7 i.e. $\pm$ 3 frames surrounding to the current

frame. The aforementioned LDA+MLLT transformations are used to reduce the dimension of feature- and model-space upto 40 and to obtain the decorrelated properties. The speaker variability from both the feature- and the model-space is normalized by employing the fMLLR-based linear regression transformation in the SAT framework. The bold face WER values and the acronyms 'NI' shown in these tables correspond to the case of best speech recognition performances and no relative improvement in the speech recognition performances compared to the conventional cases, respectively.

A severe degradation in the performances is observed for the stressed speech cases in comparison to that for the neutral speech as evident from the WERs presented in Table 4.2−Table 4.5. It has been noted that, the acoustic modelling approach using the SGMM-HMM system has resulted in the improved speech recognition performances for the stressed speech, when compared to the GMM-HMM system using both the explored speech parametrization techniques. In comparison to the conventional case, using proposed method, the performances of SGMM-HMM systems developed on the MFCC features are improved by the decrement in the WER values by 76.25%, 79.02%, 86.52% and 81.73% for recognizing angry, sad, lombard and happy speech, respectively, as depicted in Table 4.3. Similar effect is observed from the WERs given in Table 4.4 and Table 4.5 for the TEO-CB-Auto-Env features. Using proposed approach, the maximum relative decrement in WERs of 86.18%, 90.73%, 92.39% and 92.64% are obtained in comparison to the conventional cases, when SGMM-HMM systems are developed on the TEO-CB-Auto-Env features for the recognition of angry, sad, lombard and happy speech, respectively. The performances of stressed speech recognition by employing the proposed stress normalization method over the DNN-based ASR systems (DNN-HMM and DNN-SGMM) using both the MFCC and the TEO-CB-Auto-Env features are determined by varying the number of hidden layers from 1 to 6 as shown in Table 4.2−Table 4.5, respectively. The WER values depicted in these tables illustrate that, the DNN-HMM system gives much more improved stressed speech recognition performances with decreased value of WERs in comparison to the WERs obtained from the DNN-SGMM system developed on both the MFCC and the TEO-CB-Auto-Env features. The performances of speech recognition using DNN-HMM system with 1, 1, 1, 5 and 2 number of hidden layers developed on the MFCC features has resulted in the maximum recognition performances by the decrement in the WERs of 92.47%, 57.14%, 82.25%, 86.40% and 80.05% for the recognition of neutral, angry, sad, lombard and happy speech, respectively, as shown in Table 4.2. Similarly, the WERs depicted in Table 4.4 show the maximum decrement in the WERs of 86%,

Table 4.2: The recognition performances for stressed speech (WER in %) using the proposed stress normalization technique employing LPC-Based posteriorgram representation with speaker adaptation using MFCC features with respect to the GMM-HMM and the DNN-HMM systems.

Feature type	Modeling approaches		Stress class	WER (in %)		% Relative reduction
				Conventional	Proposed method	
MFCC	GMM-HMM		Neutral	0.71	0	**100**
			Angry	25.71	8.71	**66.12**
			Sad	7.41	2.19	70.44
			Lombard	6.73	1.35	79.94
			Happy	7.86	2	74.55
	DNN-HMM	1 hidden layer	Neutral	1.86	0.14	92.47
			Angry	22	9.43	57.14
			Sad	13.30	2.36	**82.25**
			Lombard	11.62	2.53	78.23
			Happy	11.14	2.29	79.44
		2 hidden layers	Neutral	1.43	0.14	90.21
			Angry	19	9.14	51.89
			Sad	11.28	2.02	82.09
			Lombard	9.26	1.68	81.86
			Happy	8.57	1.71	**80.05**
		3 hidden layers	Neutral	1.29	0.14	89.15
			Angry	18.86	8.57	54.56
			Sad	9.09	2.19	75.91
			Lombard	8.42	1.35	83.97
			Happy	8.57	1.86	78.30
		4 hidden layers	Neutral	1.29	0.29	77.52
			Angry	19.43	8.43	56.61
			Sad	10.44	2.02	80.65
			Lombard	7.41	1.35	81.78
			Happy	8.43	2	76.27
		5 hidden layers	Neutral	1.71	0.43	74.85
			Angry	18.86	8.57	54.56
			Sad	9.43	2.02	78.58
			Lombard	9.93	1.35	**86.40**
			Happy	8.43	2.14	74.61
		6 hidden layers	Neutral	1.57	0.29	81.53
			Angry	18.71	9.29	50.35
			Sad	9.93	2.19	77.94
			Lombard	9.43	1.52	83.88
			Happy	9	1.86	79.33

Table 4.3: The recognition performances for stressed speech (WER in %) using the proposed stress normalization technique employing LPC-Based posteriorgram representation with speaker adaptation using MFCC features with respect to the SGMM-HMM and the DNN-SGMM systems.

Feature type	Modeling approaches		Stress class	WER (in %)		% Relative reduction
				Conventional	Proposed method	
MFCC	SGMM-HMM		Neutral	0.14	0	**100**
			Angry	48.71	11.57	**76.25**
			Sad	24.07	5.05	**79.02**
			Lombard	26.26	3.54	**86.52**
			Happy	28.14	5.14	**81.73**
	DNN-SGMM	1 hidden layer	Neutral	5.86	3.43	41.47
			Angry	32.71	29	11.34
			Sad	28.45	21.72	23.65
			Lombard	21.72	16.16	25.60
			Happy	19.43	15.43	20.59
		2 hidden layers	Neutral	2.14	1.43	33.18
			Angry	25.43	28.43	NI
			Sad	19.53	16.67	14.64
			Lombard	14.98	12.79	14.62
			Happy	13.57	11.43	15.77
		3 hidden layers	Neutral	1.71	1.71	0
			Angry	24.29	26.43	NI
			Sad	16.33	15.99	2.10
			Lombard	16.16	12.12	25
			Happy	13.71	11	19.77
		4 hidden layers	Neutral	1.71	2	-
			Angry	23.29	25.57	NI
			Sad	14.65	16.16	NI
			Lombard	15.99	12.63	21.01
			Happy	13.86	12.71	8.30
		5 hidden layers	Neutral	2.43	1.86	23.45
			Angry	22.43	25.43	NI
			Sad	14.31	15.82	NI
			Lombard	15.15	12.79	15.58
			Happy	12.86	10.43	18.89
		6 hidden layers	Neutral	2.43	2.29	5.76
			Angry	23.57	26	NI
			Sad	15.49	16.67	NI
			Lombard	15.82	13.80	12.77
			Happy	13.86	11.71	15.51

Table 4.4: The recognition performances for stressed speech (WER in %) using the proposed stress normalization technique employing LPC-Based posteriorgram representation with speaker adaptation using TEO-CB-Auto-Env features with respect to the GMM-HMM and the DNN-HMM systems.

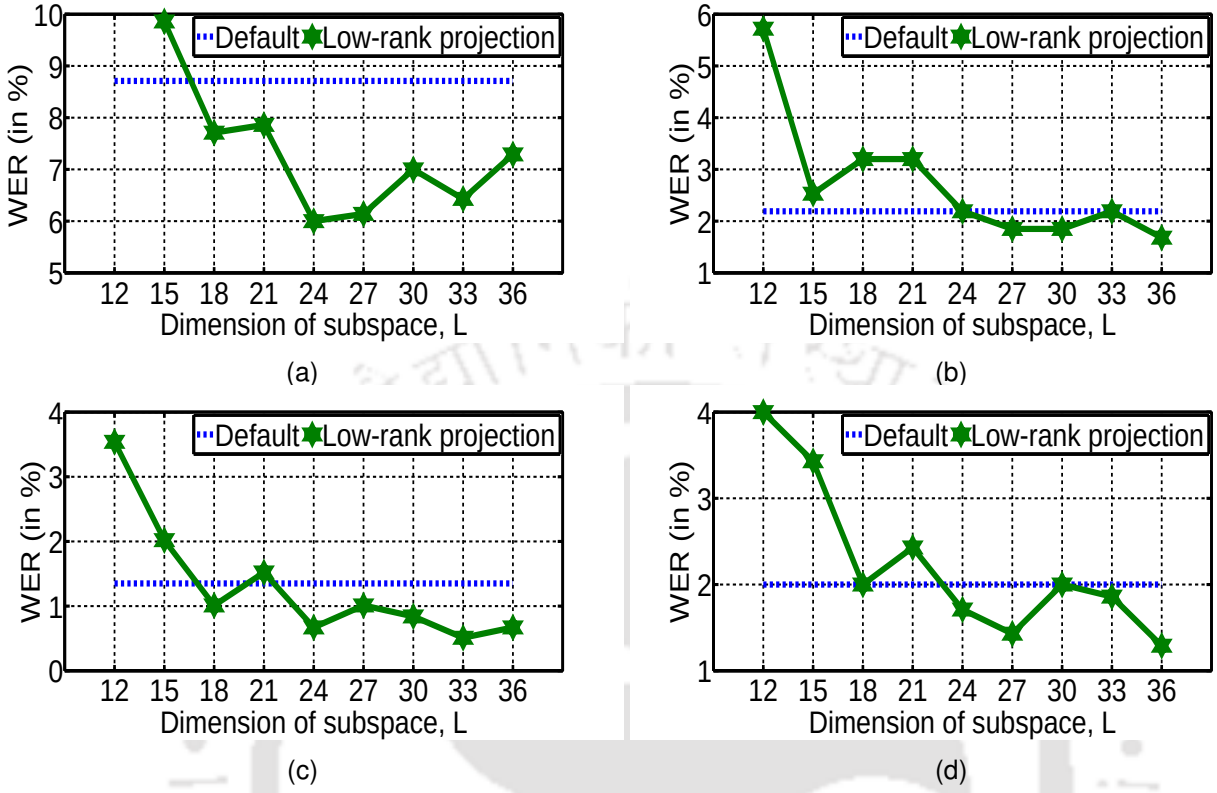
Feature type	Modeling approaches		Stress class	WER (in %)		% Relative reduction
				Conventional	Proposed method	
TEO-CB-Auto-Env	GMM-HMM		Neutral	0	0	**100**
			Angry	41.71	8.29	**80.12**
			Sad	31.31	5.72	**81.73**
			Lombard	24.07	3.37	**85.99**
			Happy	28.29	4.43	**84.34**
	DNN-HMM	1 hidden layer	Neutral	1	0.14	86
			Angry	13.29	7.14	46.27
			Sad	14.31	3.70	74.14
			Lombard	7.91	2.02	74.46
			Happy	9.57	3.57	62.69
		2 hidden layers	Neutral	0.71	0.14	80.28
			Angry	12.14	6.14	49.42
			Sad	14.48	3.20	77.90
			Lombard	7.58	2.19	71.11
			Happy	8	3.29	58.87
		3 hidden layers	Neutral	0.71	0.14	80.28
			Angry	11.29	6.71	40.57
			Sad	12.96	3.20	75.31
			Lombard	7.41	1.52	79.49
			Happy	7.29	3.43	52.95
		4 hidden layers	Neutral	0.43	0.14	67.44
			Angry	12	6.86	42.83
			Sad	12.12	3.54	70.79
			Lombard	7.74	2.36	69.51
			Happy	7.57	3.29	56.54
		5 hidden layers	Neutral	0.57	0.14	75.44
			Angry	12.43	7	43.68
			Sad	11.95	3.37	71.80
			Lombard	7.91	2.19	72.31
			Happy	8	3.29	58.87
		6 hidden layers	Neutral	0.57	0.29	49.12
			Angry	12.71	6.57	48.31
			Sad	12.12	3.20	73.60
			Lombard	8.42	2.69	68.05
			Happy	8.29	2.86	65.50

Table 4.5: The recognition performances for stressed speech (WER in %) using the proposed stress normalization technique employing LPC-Based posteriorgram representation with speaker adaptation using TEO-CB-Auto-Env features with respect to the SGMM-HMM and the DNN-SGMM systems.

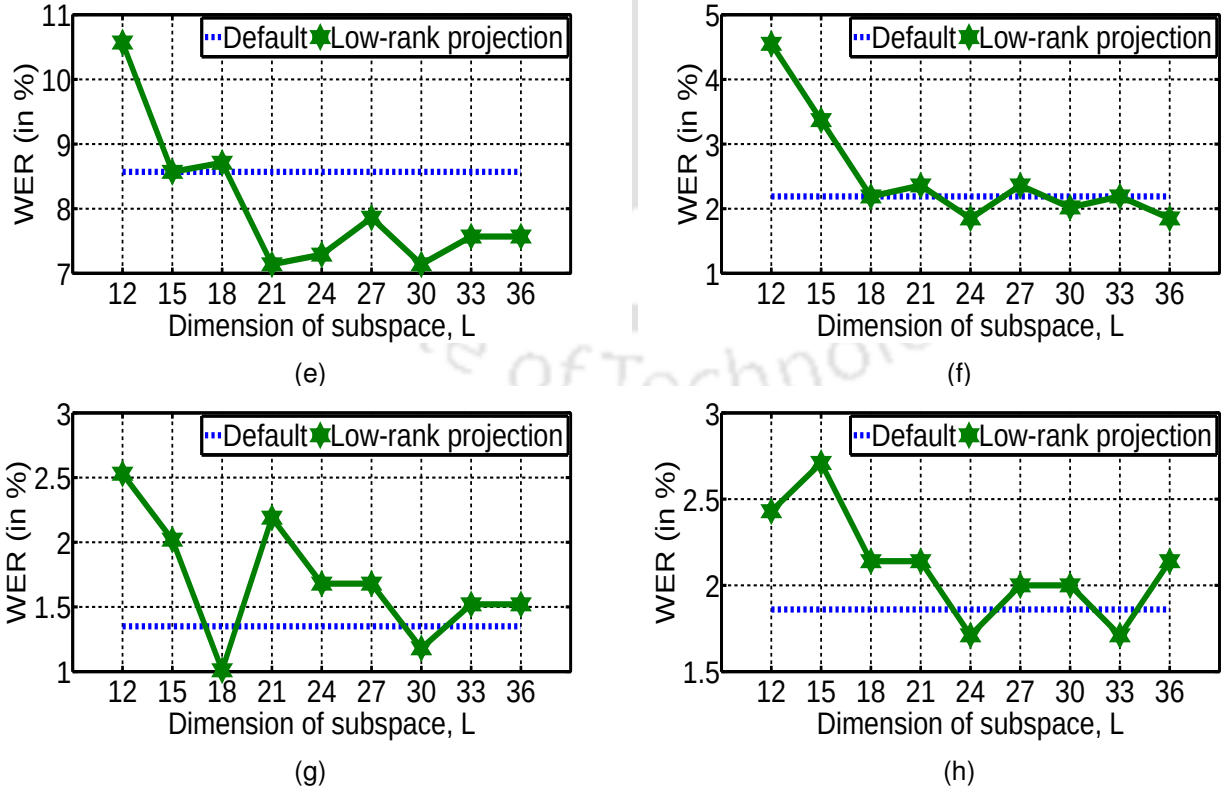
Feature type	Modeling approaches		Stress class	WER (in %)		% Relative reduction
				Conventional	Proposed method	
TEO-CB-Auto-Env	SGMM-HMM		Neutral	0	0.14	NI
			Angry	44.43	6.14	**86.18**
			Sad	36.36	3.37	**90.73**
			Lombard	28.79	2.19	**92.39**
			Happy	33	2.43	**92.64**
	DNN-SGMM	1 hidden layer	Neutral	1.29	0.43	66.67
			Angry	13.29	13.86	NI
			Sad	14.81	11.45	22.69
			Lombard	9.60	6.23	35.10
			Happy	8.86	5.86	33.86
		2 hidden layers	Neutral	0.71	0.29	59.15
			Angry	13.57	13.43	1.03
			Sad	14.31	9.93	30.61
			Lombard	7.91	5.56	29.71
			Happy	7.71	6.57	14.78
		3 hidden layers	Neutral	0.86	0.14	**83.72**
			Angry	12.57	12.43	1.11
			Sad	13.13	8.92	32.06
			Lombard	7.74	5.72	26.10
			Happy	7.57	5.43	28.27
		4 hidden layers	Neutral	0.71	0.14	80.28
			Angry	12.86	13.43	NI
			Sad	13.64	9.93	27.20
			Lombard	8.08	6.40	20.79
			Happy	8.71	6.43	26.18
		5 hidden layers	Neutral	0.86	0.14	83.72
			Angry	12	12.14	NI
			Sad	12.12	9.43	22.19
			Lombard	7.74	6.40	17.31
			Happy	8.29	5.86	29.31
		6 hidden layers	Neutral	0.86	0.14	83.72
			Angry	13.14	11.14	15.22
			Sad	13.47	9.43	30
			Lombard	8.75	6.23	28.80
			Happy	9.14	6	34.35

49.42%, 77.90%, 79.49% and 65.50% for recognizing neutral, angry, sad, lombard and happy speech over the DNN-HMM system with 1, 2, 2, 3 and 6 hidden layers trained using the TEO-CB-Auto-Env features, respectively. The performances of proposed stress normalization method have not much more affected by varying the number of hidden layers used in the DNN-HMM and the DNN-SGMM systems, except in case of high and low number of hidden layers. As discussed in Subsection 2.3.2 in Chapter 2, the deterioration in the speech recognition performances in case of using high and low number of hidden layers acknowledge the elementary and the complex non-linear mapping between the input and the output of DNN-based ASR system. These experimental observation illustrate that, the proposed subspace projection of vocal-tract system parameters for the neutral and the stressed speech onto the Gaussian-subspace consisting of vocal-tract system parameters of neutral speech utterances reduces the stress-specific attributes and preserve the acoustic characteristics. Consequently, the synthesis of neutral and stressed speech using their corresponding estimated vocal-tract system parameters reduces the variance mismatch between them. Moreover, the employed LDA-based low-rank subspace projection in the MLLT-based semi-tied adaptation technique is appeared effective for decorrelating the features of synthesized speech signals and model parameters. These decorrelated features and model parameters creates the feature- and the model-space having decorrelated characteristics, respectively. Furthermore, the fMLLR-based adaptation of feature-space in SAT mode effectively reduces the dissimilarities related to the speaker-specific attributes from the synthesized speech. The ASR system, when trained and tested using the decorrelated features of synthesized neutral and stressed speech, in which speaker variability is also suppressed exhibit the similar characteristics between the training and test environments, respectively. Therefore, the resulting ASR systems have been noted to be improved speech recognition performances in all the stress classes explored in this work.

The performance of the proposed approach for normalizing the stress-specific attributes in the lower dimensional subspace for all the explored acoustic features and modelling approaches are depicted by the WER-profiles shown in Figure 4.7−Figure 4.10. As discussed in the above paragraph, the DNN-based ASR system with the adequate number of hidden layers effectively model the different speech units of synthesized neutral and synthesized stressed speech and it has resulted in convincing improvement in the recognition performances for the stressed speech. On account of these observations, the effectiveness of the proposed stress normalization method in the lower dimensional

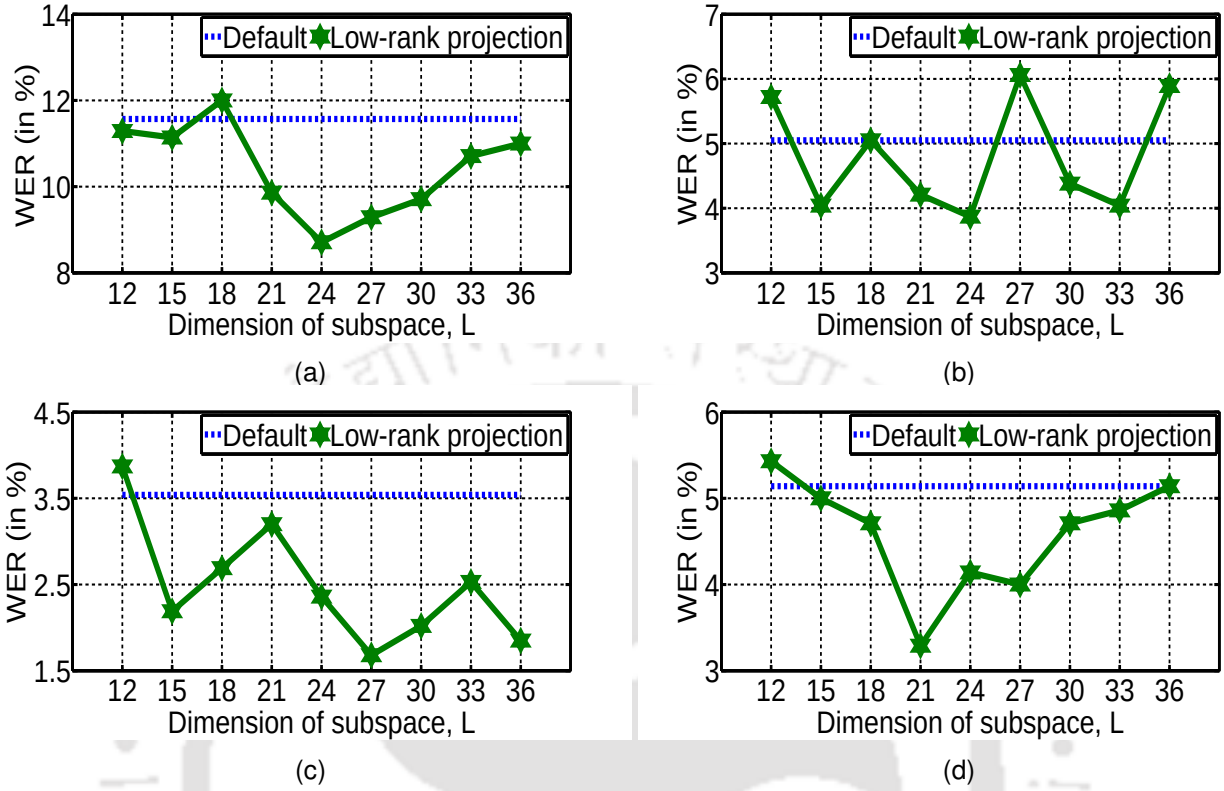


Low-rank projection over the GMM-HMM system; (a) angry, (b) sad, (c) lombard and (d) happy

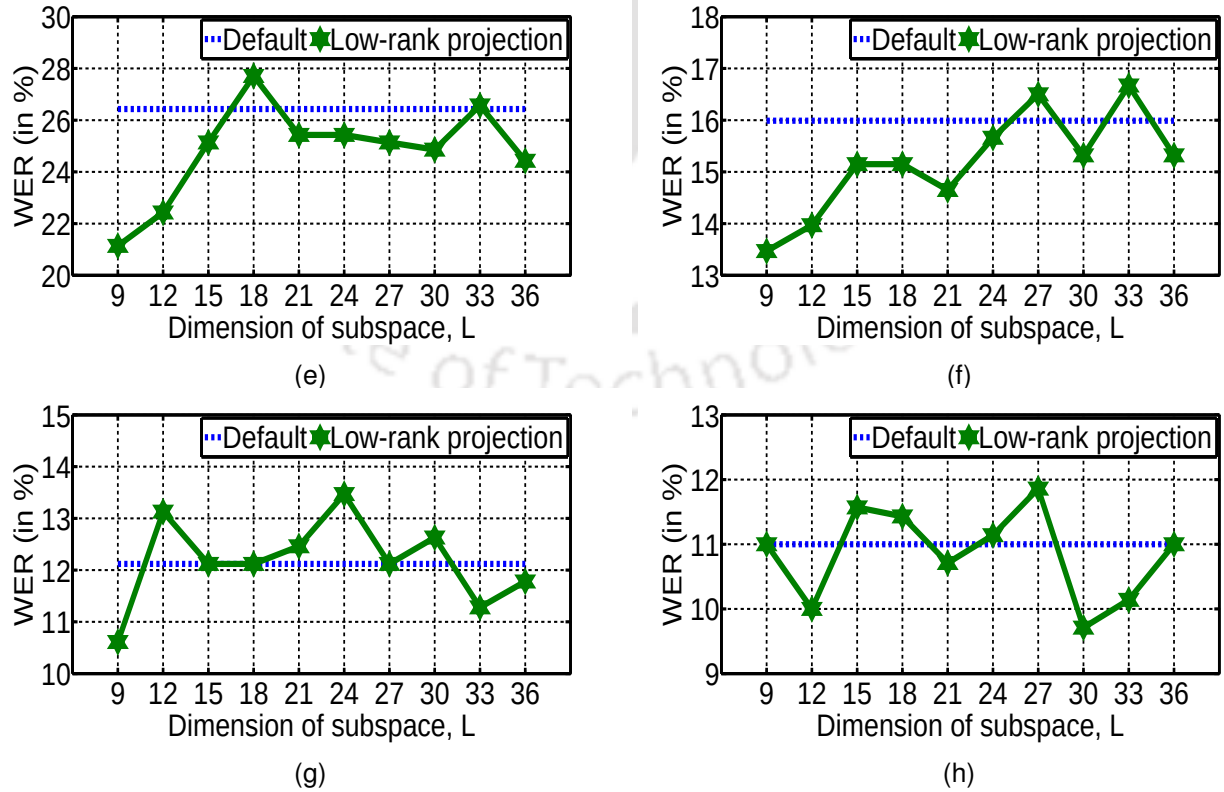


Low-rank projection over the DNN-HMM system; (e) angry, (f) sad, (g) lombard and (h) happy

Figure 4.7: Change in the WERs using the proposed stress normalization technique employing the LPC-Based posteriorgram representation with speaker adaptation method with respect to the separate GMM-HMM and DNN-HMM systems developed on the MFCC features.

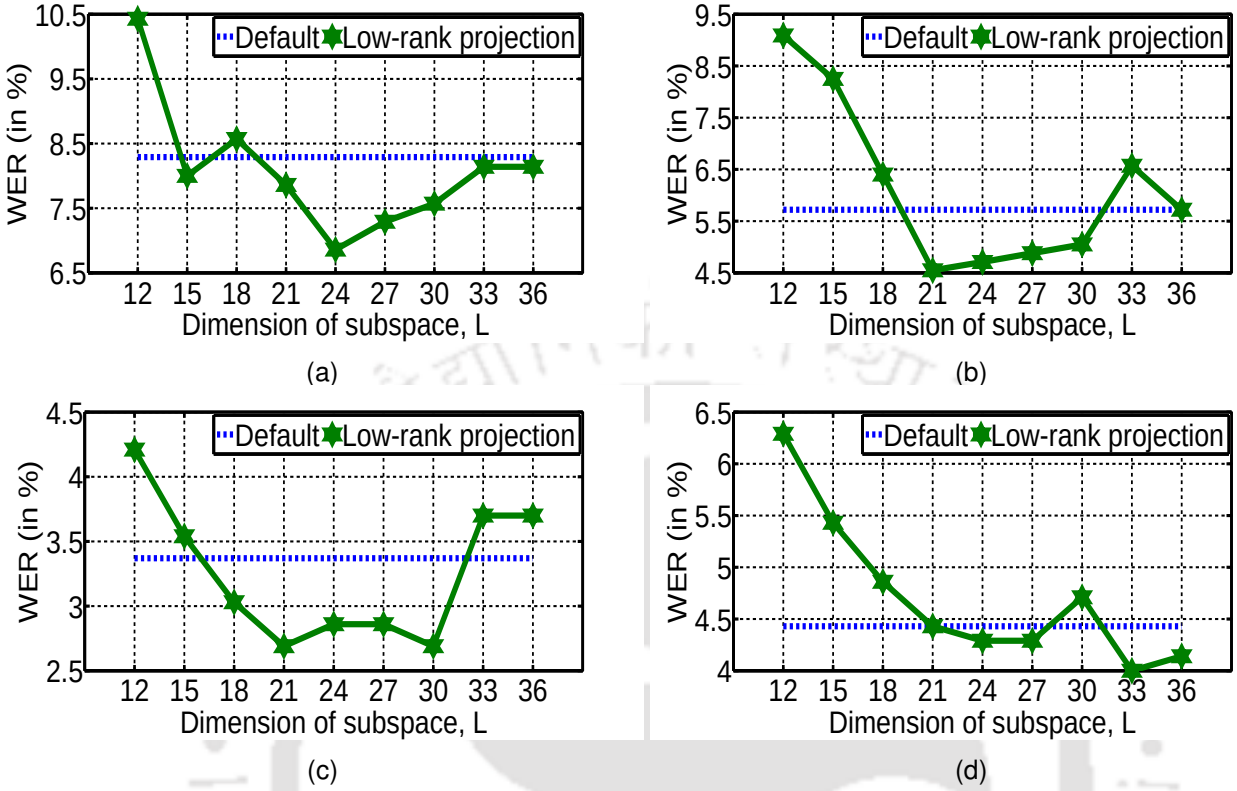


Low-rank projection over the SGMM-HMM system; (a) angry, (b) sad, (c) lombard and (d) happy

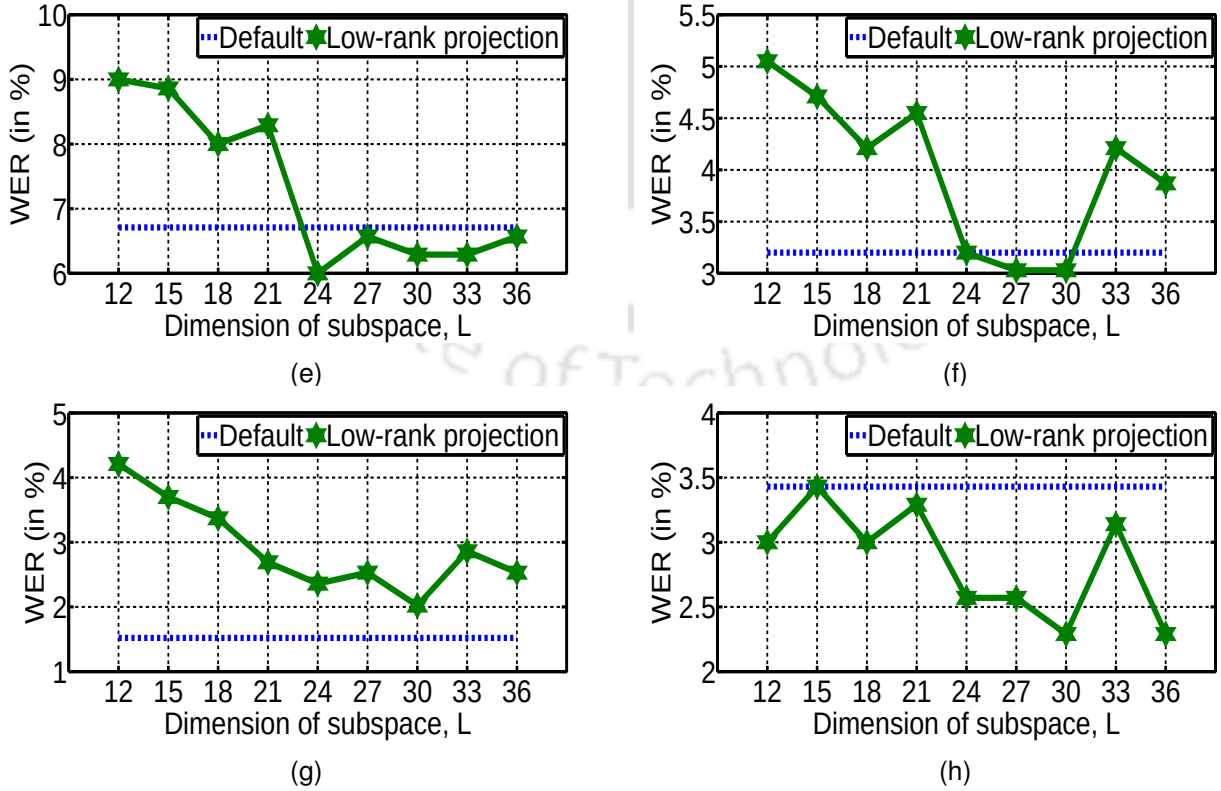


Low-rank projection over the DNN-SGMM system; (e) angry, (f) sad, (g) lombard and h) happy

Figure 4.8: Change in the WERs using the proposed stress normalization technique employing the LPC-Based posteriorgram representation with speaker adaptation method with respect to the separate SGMM-HMM and DNN-SGMM systems developed on the MFCC features.

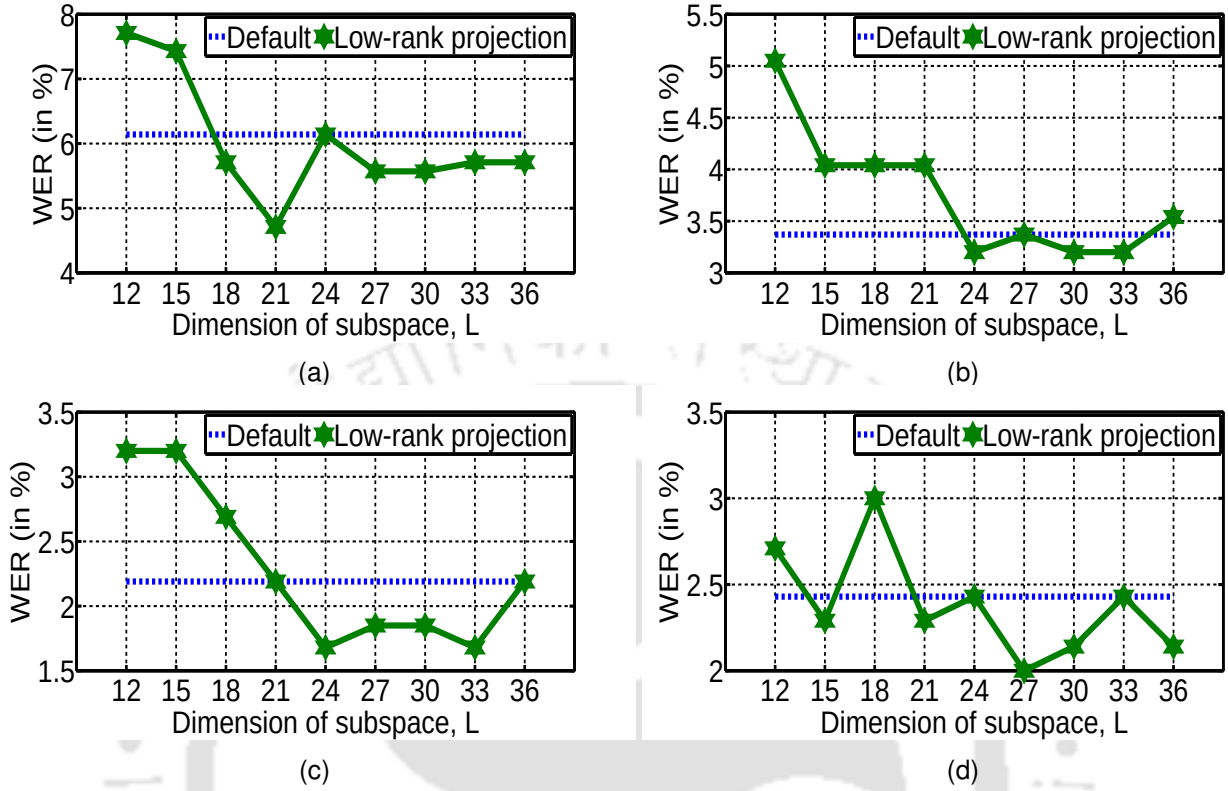


Low-rank projection over the GMM-HMM system; (a) angry, (b) sad, (c) lombard and (d) happy

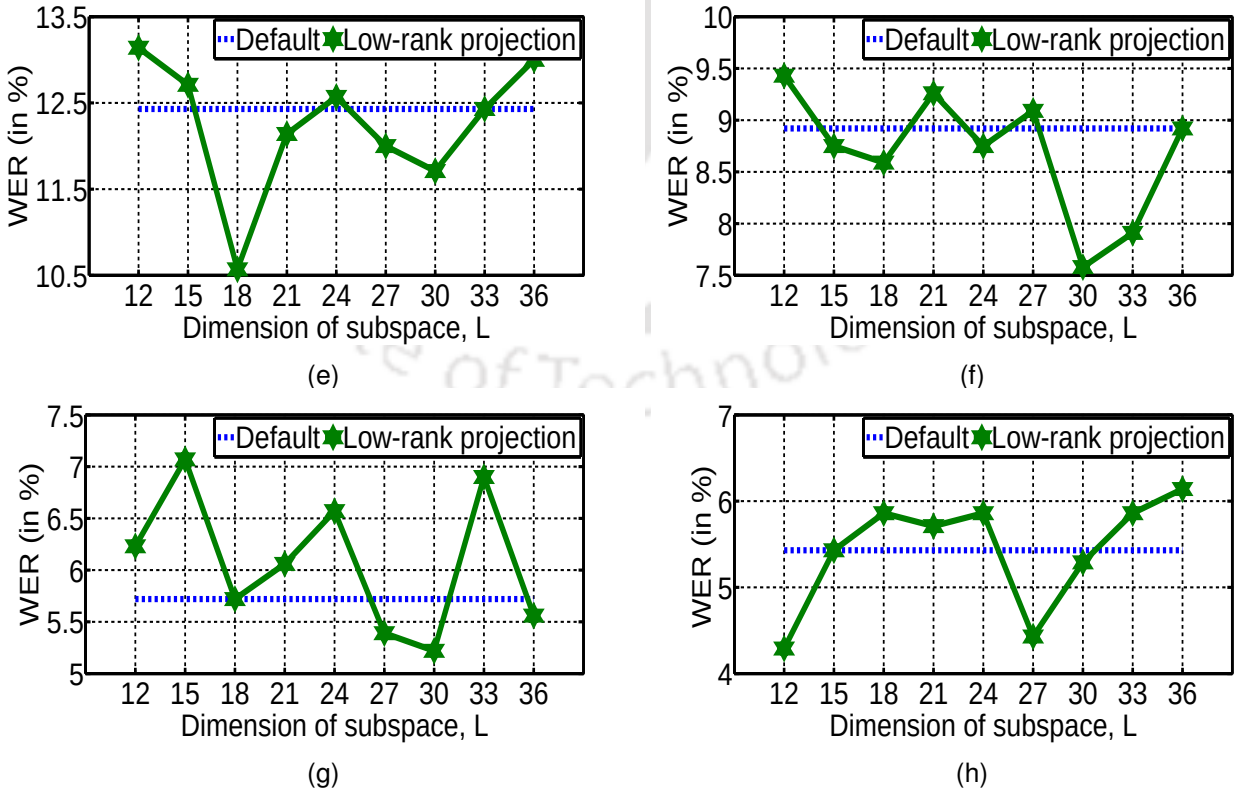


Low-rank projection over the DNN-HMM system; (e) angry, (f) sad, (g) lombard and (h) happy

Figure 4.9: Change in the WERs using the proposed stress normalization technique employing the LPC-Based posteriorgram representation with speaker adaptation method with respect to the separate GMM-HMM and DNN-HMM systems developed on the TEO-CB-Auto-Env features.



Low-rank projection over the SGMM-HMM system; (a) angry, (b) sad, (c) lombard and (d) happy



Low-rank projection over the DNN-SGMM system; (e) angry, (f) sad, (g) lombard and (h) happy

Figure 4.10: Change in the WERs using the proposed stress normalization technique employing LPC-Based posteriorgram representation with speaker adaptation method with respect to the separate SGMM-HMM and DNN-SGMM systems developed on the TEO-CB-Auto-Env features.

subspace derived by employing the LDA-based low-rank projection in MLLT-based semi-tied adaptation technique is investigated for the DNN-HMM and the DNN-SGMM system for the 3 hidden layers. In these studies, the WER-profiles corresponding to the default case are for the 40-dimensional features determined using the proposed method as discussed earlier. The rank of subspace projection matrix is varied from 36 to 12 in steps of 3, to obtain the WER-profiles for low-rank subspace projection case. For every reduction in the rank of the projection matrix, a separate ASR system is trained and tested with corresponding reduced dimensional features. This study is performed for both the MFFC and the TEO-CB-Auto-Env features. After observing these WER-profiles, it is evident that, the proposed low-rank subspace projection of the synthesized neutral and the synthesized stressed speech onto the another common subspace consisting of decorrelated characteristics, in which speaker variabilities are also normalized significantly reduces the acoustic mismatch between them. This behavior is consistently exhibited for all kinds of features and modelling paradigms explored in this work. The relative reduction in the WERs using the best performing low-rank features out of the default rank features for the MFFC as well as the TEO-CB-Auto-Env features with respect to the GMM-HMM, the SGMM-HMM, the DNN-HMM and the DNN-SGMM systems are summarized in Table 4.6 and Table 4.7, respectively. Using the proposed stress normalization method, the GMM-HMM systems trained on the MFCC features having dimension $L = 24$, $L = 33$ and $L = 36$ have resulted in the best recognition performances with the relative decrement in the WERs of 31.11%, 62.22% and 55.04% for the recognition of angry, lombard and happy speech in comparison to the WERs obtained using the default-rank features as shown in Figure 4.7(a), Figure 4.7(c) and Figure 4.7(d), respectively. Whereas, for the recognition of sad speech, the SGMM-HMM system developed on the $L = 24$ dimensional MFCC features leads to the maximum relative improvement in the recognition performances with the decrement in the WER of 23.37%, when compared to the default-rank subspace projection as depicted from the WER-profile shown in Figure 4.8(b). Similarly, using SGMM-HMM systems developed on the TEO-CB-Auto-Env features having dimension $L = 21$ and $L = 33$ features yield the maximum relative improvement in the recognition performances by the decrement in the WERs of 23.29% for the recognition of both the angry and the lombard speech in comparison to the WER values obtained in default-rank subspace projection cases as shown in Figure 4.10(a) and Figure 4.10(c), respectively. The GMM-HMM system trained using the $L = 21$ dimensional TEO-CB-Auto-Env features has resulted in the maximum relative improvement in the recognition performances

Table 4.6: Percentage of relative reductions in WERs obtained using the proposed LPC-Based posteriorgram representation with speaker adaptation method. The performances are given for the MFCC features with respect to GMM-HMM, SGMM-HMM, DNN-HMM and DNN-SGMM systems. Default corresponds to the case of using 40-dimensional feature vector while best case performances are given for the low-rank subspace projection.

Modeling approach	Stress class	WER (in %)		% Relative reduction
		Default-rank projection	Low-rank projection	
GMM-HMM	Angry	8.71	6	**31.11**
	Sad	2.19	1.68	23.29
	Lombard	1.35	0.51	**62.22**
	Happy	2	1.29	**55.04**
SGMM-HMM	Angry	11.57	8.71	24.72
	Sad	5.05	3.87	**23.37**
	Lombard	3.54	1.68	52.54
	Happy	5.14	3.29	35.99
DNN-HMM	Angry	8.57	7.14	16.69
	Sad	2.19	1.85	15.52
	Lombard	1.35	1.01	25.18
	Happy	1.86	1.71	8.06
DNN-SGMM	Angry	26.43	21.14	20.01
	Sad	15.99	13.47	15.76
	Lombard	12.12	10.61	12.46
	Happy	11	9.71	11.73

for the sad speech with the decrement in the WER by 20.45% in comparison to the WERs obtained using the default-rank features as shown in Figure 4.9(b). For the recognition of happy speech, the DNN-HMM system developed on the $L = 30$ dimensional TEO-CB-Auto-Env features yields the maximum relative improvement in the recognition performance with the decrement in the WER of 33.24% in comparison to the WER obtained in default-rank subspace projection case as depicted from the

Table 4.7: Percentage of relative reductions in WERs obtained using the proposed LPC-Based posteriorgram representation with speaker adaptation method. The performances are given for the TEO-CB-Auto-Env features with respect to GMM-HMM, SGMM-HMM, DNN-HMM and DNN-SGMM systems. Default corresponds to the case of using 40-dimensional feature vector while best case performances are given for the low-rank subspace projection.

Modeling approach	Stress class	WER (in %)		% Relative reduction
		Default-rank projection	Low-rank projection	
GMM-HMM	Angry	8.29	6.86	17.25
	Sad	5.72	4.55	**20.45**
	Lombard	3.37	2.69	20.18
	Happy	4.43	4	9.71
SGMM-HMM	Angry	6.14	4.71	**23.29**
	Sad	3.37	3.20	5.04
	Lombard	2.19	1.68	**23.29**
	Happy	2.43	2	17.70
DNN-HMM	Angry	6.71	6	10.58
	Sad	3.20	3.03	5.31
	Lombard	1.52	2.02	NI
	Happy	3.43	2.29	**33.24**
DNN-SGMM	Angry	12.43	10.57	14.96
	Sad	8.92	7.58	15.02
	Lombard	5.72	5.22	8.74
	Happy	5.43	4.29	20.99

WER-profile shown in Figure 4.9(h). These experimental results demonstrate that, the proposed low-rank subspace projection method helps in discarding the discrepancy induced due to the stress in the high-rank features. The acoustic properties of stressed speech in the low frequency region become closer to the acoustic properties of the neutral speech. Consequently, the ASR system, developed on the low-rank decorrelated features of synthesized neutral speech, in which speaker-

specific attributes are also suppressed, when tested using the corresponding reduced dimensional decorrelated features of stressed speech having normalize speaker variability exhibit the similar characteristics between the training and test environments. Therefore, the resulting speaker dependent ASR systems have been noted in consistent improvement in the recognition performances for the stressed speech in all the stress classes studied in this work. The proposed LPC-based posteriorgram representation with speaker adaptation technique effectively reduces the acoustic mismatch between the neutral and the stressed speech in the lower dimensional subspace.

4.6 Summary

The work presented in this chapter is intended towards the normalization of stress information by investigating the modification in the vocal-tract system characteristics under stress condition. A novel subspace projection-based approach over the vocal-tract system parameters has been explored to develop an effective stress normalization technique. The characteristics of vocal-tract system are captured using the linear prediction coefficients (LPCs). To reduce the acoustic mismatch between the neutral and the stressed speech, their corresponding LPCs are projected onto a common Gaussian-subspace and it has resulted in the posteriorgram features. Our study shows that, the creation of this Gaussian-subspace using the LPCs of the neutral speech utterances estimates the LPCs for the stressed speech with the characteristics similar to the LPCs for the neutral speech. The experimental evaluations using the visual and the error analysis demonstrate that, the synthesis of neutral and stressed speech by employing the posteriorgram features with respect to their corresponding estimated LPCs significantly reduces the variance mismatch between them. Consequently, the ASR system, when trained and tested using the synthesized neutral and the synthesized stressed speech exhibits the similar acoustic characteristics for training and test environments, respectively. The resulting ASR system has been noted to be improved stressed speech recognition performances with decreased values of WER in comparison to the WERs obtained using the conventional case for all the studied stress classes.

The next issue we take up is that of the speaker adaptation technique for increasing the robustness of the proposed stress normalization method. The fMLLR-based transformation is employed in the SAT framework and is reported to be very effective in reducing the speaker-specific variabilities from the synthesized speech. In this Chapter, we have also seen the usefulness of proposed stress

normalization method in the lower dimensional subspace. The LDA-based low-rank subspace projection in MLLT-based semi-tied adaptation technique developed on the synthesized neutral speech has been found to be very efficacious for the projection of feature- and model-space onto the decorrelated subspace of lower dimension. Experimental evaluations presented in this work reveal that, the projection of the synthesized speech onto a lower dimensional subspace leads to the significant improvements in recognition performances of stressed speech of all the explored stress classes on the GMM-, the SGMM-based ASR system but also on the DNN-based modelling cases.

5

Sparse Representation of LPC Over Utterance-Specific Adaptive Dictionary for Stress Normalization

Contents

5. Sparse Representation of LPC Over Utterance-Specific Adaptive Dictionary for Stress Normalization

In this final Chapter, we are concerned with the problem of efficiently representing the vocal-tract system parameters for the development of effective stress normalization technique. The experimental evaluations presented in Chapter 4 illustrate that, the analysis of modifications in the vocal-tract system under stress condition has led to an fascinating approach for reducing the variance mismatch between the neutral and the stressed speech. In those studies, the subspace projection was derived using the full-rank subspace projection matrix by exploiting the posteriorgram representation of vocal-tract system parameters and appeared very effective for normalizing the stress-specific attributes. These facts motivate us for further investigation on the changes in the vocal-tract system for the stressed speech. The investigation on the modification in the vocal-tract system by exploring the subspace projection technique using the over-complete subspace projection matrix in sparse domain, which is generally referred to as the sparse representation can help in designing the effective algorithm for reducing the acoustic mismatch between the different speech units of neutral and stressed speech. It is hypothesized that, the synthesis of neutral and stressed speech utterances by employing their corresponding estimated vocal-tract system parameters can result in the speech signals with the characteristics similar to the neutral speech.

The subspace modelling using the sparse representation has attracted the interest of many researchers in the field of stressed speech processing [42, 47, 51, 74, 75, 165–170]. The sparse representation is a powerful technique for the reconstruction and the compression of signals [76–78]. In this work, the acoustic mismatch between the vocal-tract system parameters for the neutral and the stressed speech has been reduced by introducing the sparse representation technique through the linear transformation on the over-complete linear models also called as the dictionary under which the observed signal can be sparsely coded using a few suitable bases. The sparse representation of vocal-tract system parameters provides its decomposition using a few predefined atoms of dictionary. The robust sparse representation depends on the estimation of an effective dictionary. The creation of an effective dictionary incorporates the following two factors: the learning mechanism and the selection of input data, respectively. The efficient learning mechanism and the informative selection of input data will lead to the estimation of an effective dictionary. This can yield to the best linear combination over the dictionary atoms under sparsity for each element of input data. The estimation of dictionary using the acoustic parameters of vocal-tract system under neutral condition will significantly alleviate the impact of stress as well as retain the acoustic information, since it con-

sists of acoustically rich phonetic information. In this Chapter, the K-SVD algorithm [76] is employed to create the effective dictionary. The K-SVD algorithm comprises the jointly update of dictionary atoms along with update of sparse coding. Consequently, the learning of dictionary by exploiting this iterative process using the vocal-tract system parameters of neutral speech utterances can help in developing the effective dictionary. This dictionary comprises of a fixed number of atoms and called as the global dictionary. Moreover, in literature, numerous studies have been reported on the alteration in the duration of speech signals produced under stress conditions [5, 6, 11, 60, 70]. In those studies, the considerable degree of modification in the duration parameter was noted for the production of word, vowel, semi-vowel, consonant, diphthong etc. and were found to be very effective for increasing the performances of tasks involving the recognition and the classification of stressed speech. Therefore, the sparse representation of vocal-tract system parameters of neutral or stressed speech utterance through the linear combination over dictionary, in which the number of atoms depends on the information about the duration parameter of that speech utterance can significantly reduce the aforementioned acoustic mismatch. In order to address this, the K-nearest-neighbour (K-NN) algorithm-based non-parametric probability density estimation method [143, 147] has been introduced to incorporate the information about the duration parameter of speech utterance in the estimation of an effective dictionary. The dimension of this exemplar dictionary varies according to the duration of speech signal and is referred to as the utterance-specific adaptive dictionary. Both the proposed invariable size global dictionary and utterance-specific adaptive dictionary have been reported very effective for reducing the acoustic mismatch between the vocal-tract system parameters of neutral and stressed speech.

In Chapter 4, the normalization of speaker-specific attributes is found to be very effective in increasing the performance of proposed stress normalization method with the significant improvement in the speech recognition performances for the stressed speech. Motivated by the same, in this Chapter, we have further intended to explore the speaker normalization to increase the effectiveness of the proposed approach employing the sparse representation of vocal-tract system parameters for normalizing the stress information. The effectiveness of the proposed stress normalization method is evaluated using the two sets of database namely: the Speech under Simulated Stress Condition (SUSSC) database [118] and the Database of German Emotional Speech (Emo-DB) [119], respectively. The experimental evaluations have been accomplished by analyzing the waveform, the spectral

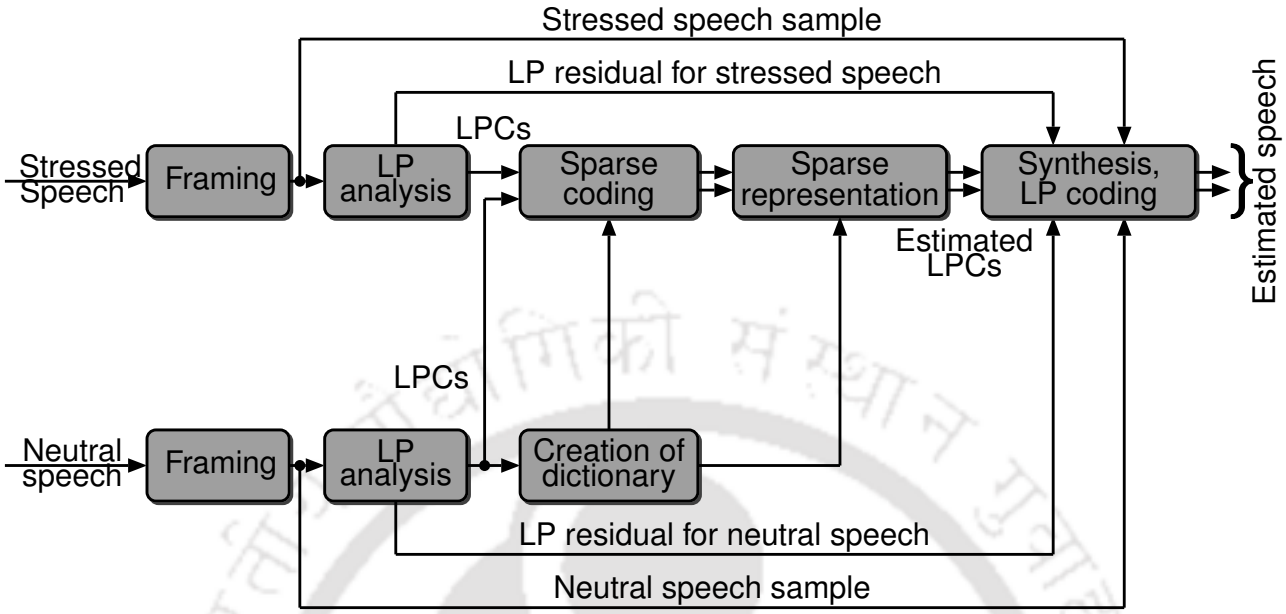


Figure 5.1: Stress normalization using the proposed synthesis process employing the sparse representation of vocal-tract system parameters.

distribution, the Kullback Leibler (KL) divergence [140, 141] and the speech recognition for measuring the effectiveness of the proposed stress normalization technique.

The remainder of this Chapter is organized as follows: The proposed stress normalization method employing the sparse representation of vocal-tract system parameters is presented in Section 5.1. In Section 5.2, the synthesized speech signals are quantified. The methodology of speaker normalization using the semi-tied adaptation technique is summarized in Section 5.3. Section 5.4 contains the descriptions on the experimental setup used followed by the results and discussions of the proposed stress normalization technique. Finally, the findings of this Chapter are summarized in Section 5.5.

5.1 Sparse Representation of Vocal-Tract System Parameters for Speech Synthesis

Stressed speech differs considerably from neutral speech in many important aspects and characteristics, thus resulting in a high variance mismatch. To reduce the variance mismatch, in this Chapter, we have introduced the sparse modelling of vocal-tract system parameters. The parameters of vocal-tract system are modeled using the linear prediction coefficients (LPCs), which are extracted by processing the speech signals through the linear prediction (LP) analysis [1, 101–105, 127]. The

steps involved in the proposed stress normalization method is described in the block diagram shown in Figure 5.1. It is evident from this block diagram that, the proposed stress normalization technique comprises mainly of two-stages: the synthesis and the estimation of an effective dictionary, respectively. The methodology for the synthesis of neutral and stressed speech using the vocal-tract system parameters, which comprise the normalize stress information are similar to as described in Subsection 4.1.1 in Chapter 4. The synthesized neutral and the synthesized stressed speech are considered as the speech constituting the similar acoustic properties and is computed using Eq. (4.1) and Eq. (4.2) given as, $\tilde{s}_f(n) = e_f(n) - \sum_{p=1}^{P} \tilde{a}_{f_p} s_f(n-p)$ and $e_f(n) = s_f(n) + \sum_{p=1}^{P} a_{f_p} s_f(n-p)$. Where, $\{a_{f_p}\}_{p=1}^{P}$ and $e_f(n)$ are the LPCs and the LP residual error of the speech sample $s_f(n)$, derived using the P-order LP analysis. The estimated LPCs are represented by $\{\tilde{a}_{f_p}\}_{p=1}^{P}$ and are considered as the LPCs having normalize stress information. It is observed that, the effectiveness of the proposed stress normalization method depends on an effective estimation of $\{\tilde{a}_{f_p}\}_{p=1}^{P}$. The proposed stress normalization technique explores the sparse representation of $\{a_{f_p}\}_{p=1}^{P}$ for the estimation of substantial $\{\tilde{a}_{f_p}\}_{p=1}^{P}$. Therefore, the learning of an effective dictionary will help in boosting the effectiveness of the proposed approach for normalizing the stress information. The following Subsection describe the estimation of vocal-tract system parameters by exploiting the sparse representation technique. This is followed by details of the methodology of estimation for the effective dictionary.

5.1.1 Sparse Representation of Vocal-Tract System Parameters

In this section, we describe the method of estimation of vocal-tract system parameters for normalizing the stress information. The proposed stress normalization technique incorporates the synthesis process. Therefore, the estimation of vocal-tract system parameters for the neutral and the stressed speech, which constitute the similar acoustic characteristics can help in normalizing the stress information. In this work, the sparse representation technique has been explored for the estimation of vocal-tract system parameters with the characteristics similar to the vocal-tract system parameters for the neutral speech. The proposed approach for the estimation of vocal-tract system parameters are described in Figure 5.2. As discussed earlier, neutral speech constitutes the acoustically rich phonetic information. Therefore, the estimation of dictionary using the neutral speech data helps in normalizing the stress-specific divergences as well as in capturing the phonetic information. The following steps summarize the proposed method for the estimation of vocal-tract system parameters.

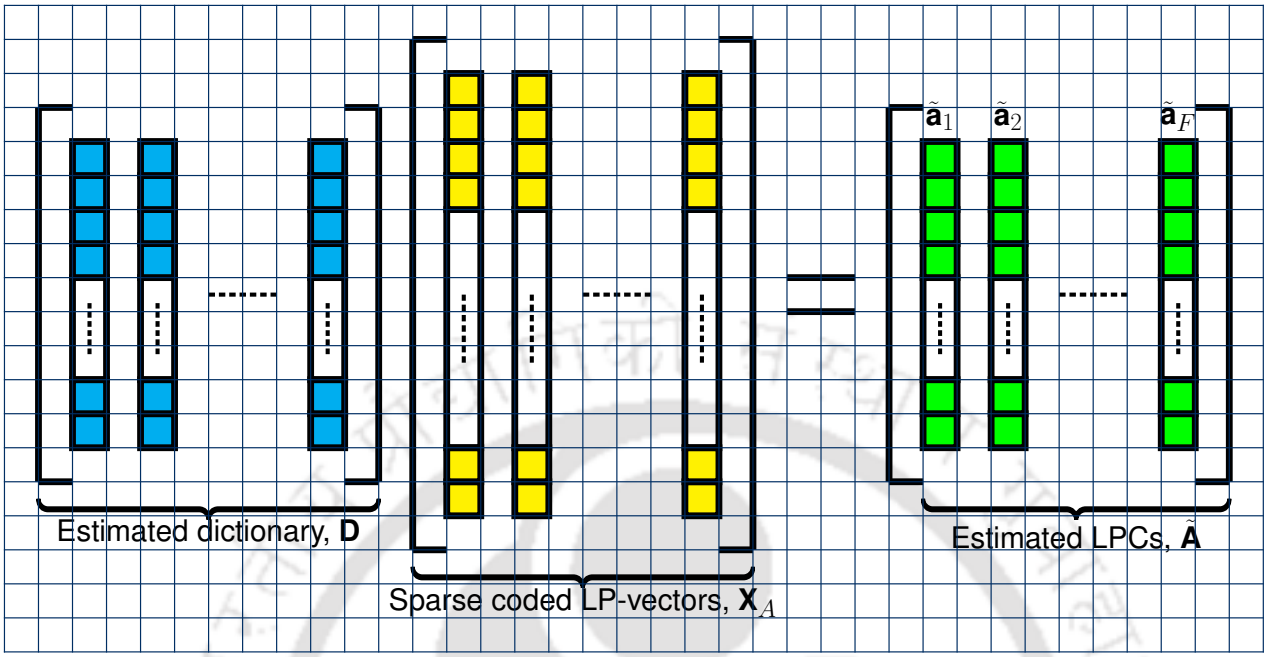


Figure 5.2: Sparse representation of LPCs for the normalization of stress-specific divergences.

Step I: At first, the neutral speech utterances belonging to the training database are processed through the P-order LP analysis to extract the LPCs for each frame and are arranged in $\mathbf{a}_f = \begin{bmatrix} a_{f_1} & a_{f_2} & \cdots & a_{f_P} \end{bmatrix}^T$. In this work, the set of LPCs corresponding to each frame is called as the LP-vector. The dimension of the LP-vector become $P \times 1$. The superscript T and the subscript f denote the transpose operation and the frame index, respectively. In this way, the set of LP-vectors $\{\mathbf{a}_f\}_{f=1}^{N}$ corresponding to all frames of neutral speech utterances are determined and arranged in $\mathbf{Y} = \begin{bmatrix} \mathbf{a}_1 & \mathbf{a}_2 & \cdots & \mathbf{a}_N \end{bmatrix}$. Where, N represents the total number of frames of neutral speech utterances of training database.

Step II: In the next step, the dictionary **D** is constructed by employing **Y** to capture the principal dimensions of the acoustic variations represented by the vocal-tract system parameters of neutral speech utterances using the technique described in the succeeding Subsection 5.1.2.

Step III: For any given observed neutral or stressed speech utterance having F, $(1 \leq f \leq F)$ frames, the set of LP-vectors are determined and arranged in $\mathbf{A} = \begin{bmatrix} \mathbf{a}_1 & \mathbf{a}_2 & \cdots & \mathbf{a}_F \end{bmatrix}$. The set of LP-vectors corresponding to these LP-vectors in **A** are estimated by the sparse representa-

tion over the dictionary **D** as follows,

$$\tilde{\mathbf{A}} = \mathbf{D}\mathbf{X}_A \tag{5.1}$$

where, $\tilde{\mathbf{A}} = \begin{bmatrix} \tilde{\mathbf{a}}_1 & \tilde{\mathbf{a}}_2 & \cdots & \tilde{\mathbf{a}}_F \end{bmatrix}$ contains the set of estimated LP-vectors for the given neutral/ stressed speech utterance as shown in Figure 5.2. The sparse coded vectors corresponding to the LP-vectors in **A** are arranged in $\mathbf{X}_A$. The sparsity constraint is denoted by T_0. The estimated LPCs in $\tilde{\mathbf{A}}$ are considered as the LPCs having normalize stress information.

Both the neutral and the stressed speech utterances are synthesized using the elements of the corresponding estimated LP-vectors $\{\tilde{\mathbf{a}}_f\}_{f=1}^{F}$ in $\tilde{\mathbf{A}}$ as described earlier. It is assumed that, the synthesized neutral and synthesized stressed speech consists of similar acoustic characteristics.

5.1.2 Proposed Learning Mechanism of Effective Dictionary

The effectiveness of sparse representation depends on the promising creation of dictionary **D** as illustrated in Eq. (5.1). In this work, two different learning mechanisms are proposed for an estimation of effective dictionary. The first learning mechanisms exploits the well known K-SVD algorithm for the creation of invariable size global dictionary. In addition to this, we have exploited the information about the duration parameter of speech utterance to construct the exemplar dictionary. The incorporation of information about the duration parameter can be interpreted as the sparse representation over the utterance-specific adaptive dictionary is also described in the following.

Creation of Invariable Size Global Dictionary: The learning mechanism of invariable size global dictionary incorporates the K-SVD algorithm. The K-SVD, a generalization of K-means clustering method is most widely algorithm used to learn the redundant dictionary for sparse representation [76]. As discussed earlier, the sparse representation of LPCs of neutral and stressed speech utterances over the dictionary, which comprises the LPCs of neutral speech will help in reducing the acoustic mismatch between them. Therefore, the set of LP-vectors $\{\mathbf{a}_f\}_{f=1}^{N}$ of neutral speech utterances arranged in **Y** are used to estimate the dictionary as depicted in block diagram shown in Figure 5.3. The K-SVD algorithm constructs the dictionary for best sparse representation of the training vectors (LP-vectors in **Y**) with a minimum sparsity constraint using following objective function,

$$\min_{\mathbf{D},\mathbf{X}_Y} \{\| \mathbf{Y} - \mathbf{D}\mathbf{X}_Y \|_2^2\} \quad \text{subject to } \| \mathbf{x}_f \|_0 \leq T_0 \quad \forall f \tag{5.2}$$

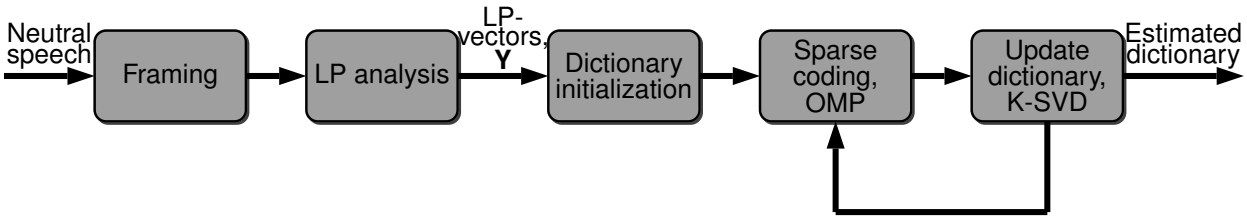


Figure 5.3: Estimation of invariable size global dictionary using the K-SVD algorithm.

where, the dictionary **D** constitutes D atoms. The dimension of this dictionary is $P \times D$. The matrix, $\mathbf{X}_Y$ having dimension $D \times N$ comprises the set of sparse coded vectors corresponding to **Y** with constraint on sparsity T_0. The learning process of dictionary involves jointly update of dictionary atoms along with update of sparse coding, thus resulting in accelerated convergence. The sparse coding or atom decomposition can be done using any of the pursuit algorithm. In this work, the orthogonal matching pursuit (OMP) [76, 78, 171–173] is used for sparse coding as depicted in Figure 5.3. The sparse representation for all the observed neutral and stressed speech utterances are derived over these invariable number of atoms of the same estimated dictionary **D**. In this work, this estimated dictionary **D** is referred to as the invariable size global dictionary due to it global exploration for sparse representation for all the observed neutral and stressed speech utterances.

Creation of Utterance-Specific Adaptive dictionary: The speaker changes the speaking rate to manifest the information about the stress condition as well to retain the acoustic information [5, 11]. Consequently, the duration of speech signal is influenced and it leads to the amended prosodic characteristics for the stressed speech in comparison to the neutral speech. The literature reviews summarized in Chapter 1 have demonstrated that, the same speech utterance produced under different stress conditions exhibit the different time duration and is found to be very informative for characterizing the stress classes. The investigation on the changes in the duration parameter is increasingly attractive courtesy from a broader range of researchers involved in the area of stressed speech due to it potentiality of classifying the different stress classes [6, 60]. In the work reported in [70], the significant degree of changes have been observed for fricatives, diphthongs, nasals, semi-vowels, affricates, stops, silence and vowels to emphasize the lombard effect. These studies show that, the duration of speech signal carries the significant information of the stressed speech. The incorporation of information about the duration parameter of speech utterance (neutral or stressed speech) in the

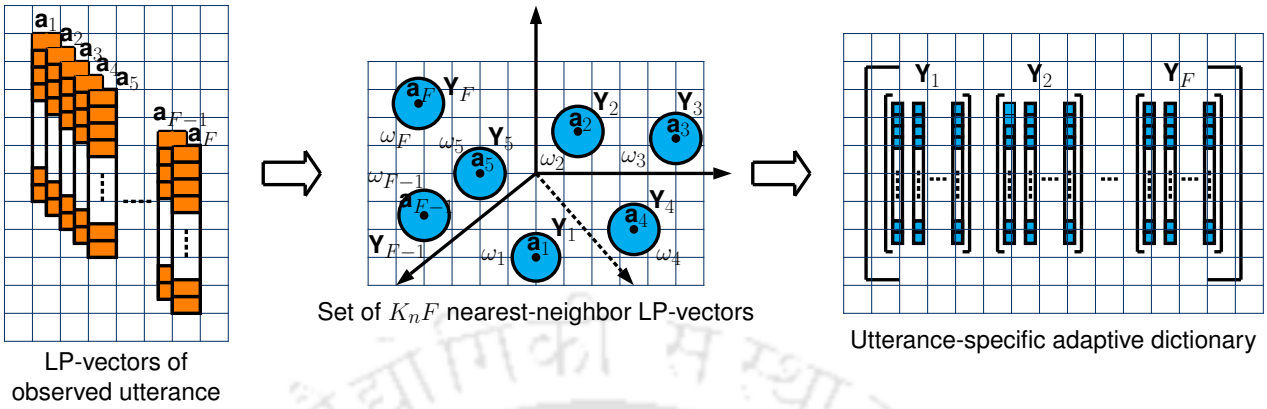


Figure 5.4: Creation of utterance-specific adaptive dictionary using the K-NN algorithm.

estimation of dictionary can help in improving the sparse representation of set of LP-vectors for that neutral/ stressed speech utterance to reduce the aforementioned acoustic mismatch. To model the information about the duration parameter of speech, we have explored the K-nearest-neighbour (K-NN) algorithm-based non-parametric probability density estimation method as shown in Figure 5.4. The K-NN algorithm is based on the conception that the probability density at a point can be estimated by growing the radius of the sphere around that point until it contains the precisely fixed number of sample observations [143, 147]. The integration of information about the duration parameter of each observed neutral or stressed speech utterance in the estimation of dictionary leads to the estimation of an adaptive dictionary according to the observed utterance. In this work, this dictionary is called as the utterance-specific adaptive dictionary. The number of atoms in this dictionary depends on the duration of observed speech utterance as depicted in Figure 5.4. The following steps describe the learning mechanism of proposed utterance-specific adaptive dictionary.

Step I: At first, LP-vectors $\{\mathbf{a}_f\}_{f=1}^{F}$ for the observed speech utterance (neutral or stressed speech) are determined using the P-order LP analysis. The subscript f represents the frame index and varies from $1 \leq f \leq F$, where F is the total frames of observed speech utterance.

Step II: In the next step, for the F number of LP-vectors $\{\mathbf{a}_f\}_{f=1}^{F}$ of observed speech utterance, the feature-space with respect to the LP-vectors of neutral speech utterances arranged in $\mathbf{Y}$ is partitioned into the F number of distinct cells $\{\omega_f\}_{f=1}^{F}$ using the Bayes decision rule by exploring the K-NN algorithm as shown in Figure 5.4. This feature-space comprises the vocal-tract system parameters of neutral speech signals. The individual cell consists of fixed K_n number

of nearest-neighbor LP-vector of neutral speech utterances corresponding to distinct LP-vector of observed utterance and generally referred to as the Voronoi tesselation of space [143, 147].

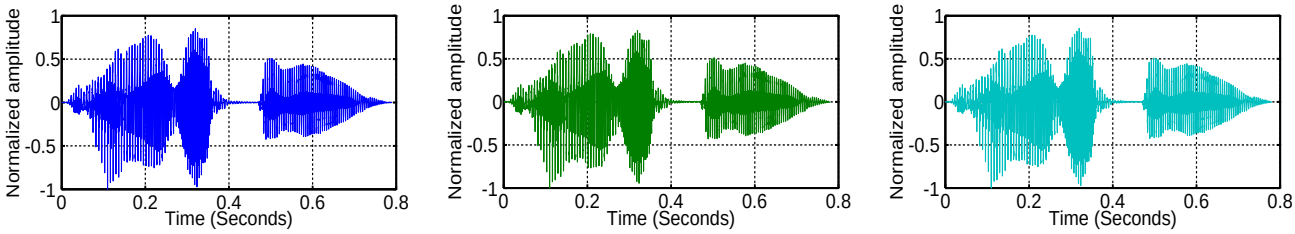
Step III: The K_n nearest-neighbor LP-vectors corresponding to each cells are arranged in $\mathbf{Y}_f = \begin{bmatrix} \mathbf{a}_{f_1} & \mathbf{a}_{f_2} & \cdots & \mathbf{a}_{f_{K_n}} \end{bmatrix}$. In this way, the set of nearest-neighbor LP-vectors $\{\mathbf{Y}_f\}_{f=1}^{F}$ are determined from all the partitioned cells. These nearest-neighbor LP-vectors in $\{\mathbf{Y}_f\}_{f=1}^{F}$ signify the estimate of the probability density corresponding to the set of LP-vectors $\{\mathbf{a}_f\}_{f=1}^{F}$ of observed speech utterance, respectively. Each column of these matrices represent one of the LP-vectors in the $\mathbf{Y}$, which comprises the vocal-tract system parameters of neutral speech signals. The dictionary is created by the concatenation of the individual matrices $\{\mathbf{Y}_f\}_{f=1}^{F}$ in $\mathbf{D} = \begin{bmatrix} \mathbf{Y}_1 & \mathbf{Y}_2 & \cdots & \mathbf{Y}_F \end{bmatrix}$. The number of atoms, K_nF in this exemplar dictionary are associated with the duration of observed speech utterance in terms of number of frames F, which makes it adaptive to the observed utterance and referred to as the utterance-specific adaptive dictionary.

The estimated utterance-specific adaptive dictionary constitutes the nearest-neighbor LP-vectors of neutral speech utterances. Therefore, the sparse representation of LP-vectors of the given observed neutral and stressed speech utterances using their corresponding estimated utterance-specific adaptive dictionaries in Eq. (5.1) will help in reducing the acoustic mismatch between them.

5.2 Quantification of Synthesized Speech

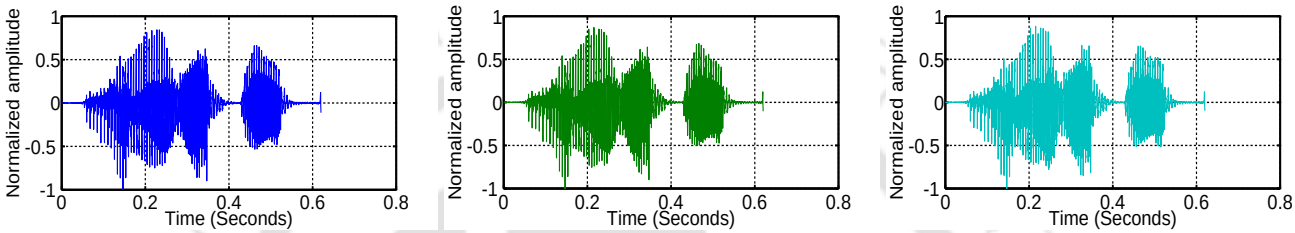
In this Section, the proposed synthesis process employing the sparse representation of vocal-tract system parameters for normalizing the stress information is quantified. The synthesized neutral and the synthesized stressed speech are considered as the speech signals constituting the similar acoustic characteristics. The quality of synthesized speech reflects the effectiveness of the proposed stress normalization technique. To measure the quality of synthesized speech, we have further explored the visual analysis and the error analysis as discussed in Section 4.2 in Chapter 4. The following paragraphs describe the quantification of synthesized speech signals by exploring the aforementioned two different but interconnected analysis using the two stressed speech databases namely: the SUSSC database [118] and the Emo-DB database [119], respectively.

Visual Analysis: In the visual analysis, we have quantified the synthesized speech signals by analyzing their characteristics in the time and the frequency domains. The time domain interpretation



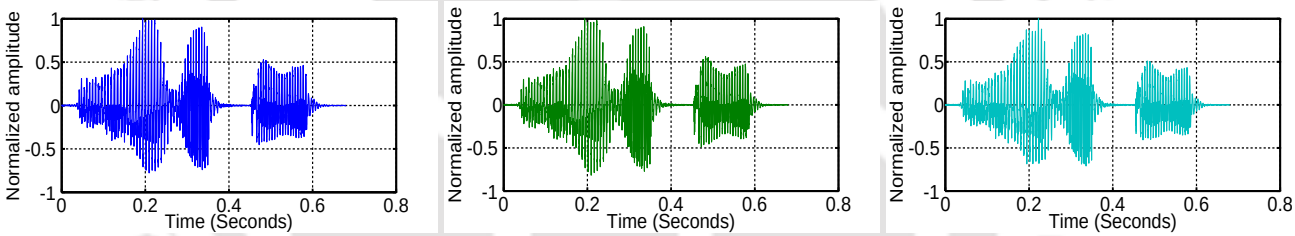
(a) (b) (c)

Speech recorded under neutral condition: (a) original (b) global dictionary and (c) adaptive dictionary



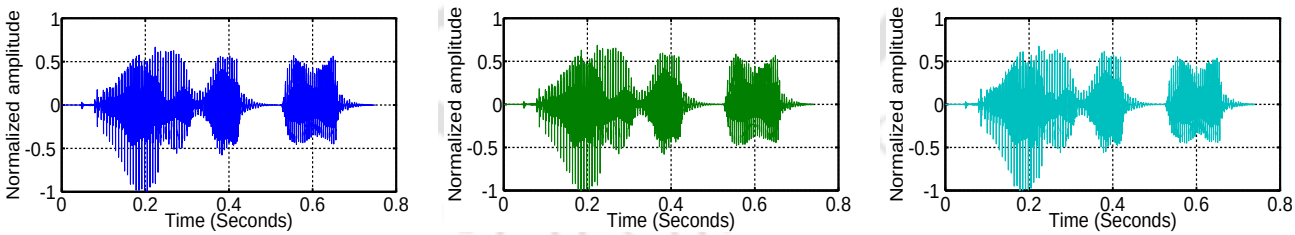
(d) (e) (f)

Speech recorded under angry condition: (d) original (e) global dictionary and (f) adaptive dictionary



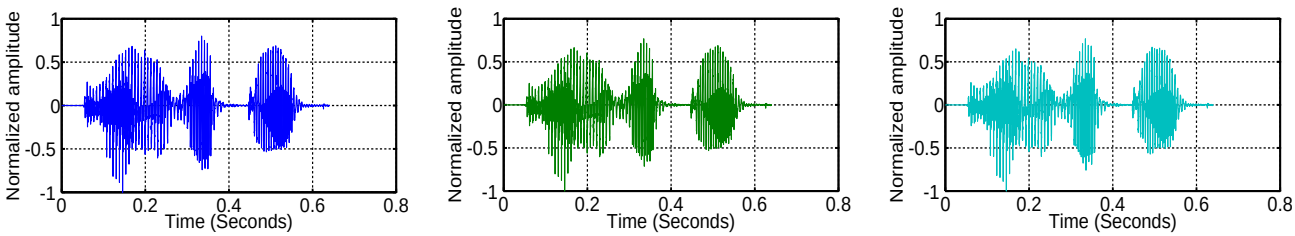
(g) (h) (i)

Speech recorded under sad condition: (g) original (h) global dictionary and (i) adaptive dictionary



(j) (k) (l)

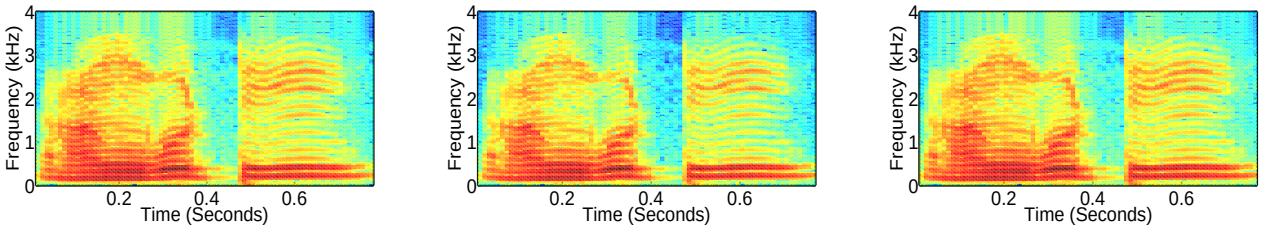
Speech recorded under lombard condition: (j) original (k) global dictionary and (l) adaptive dictionary



(m) (n) (o)

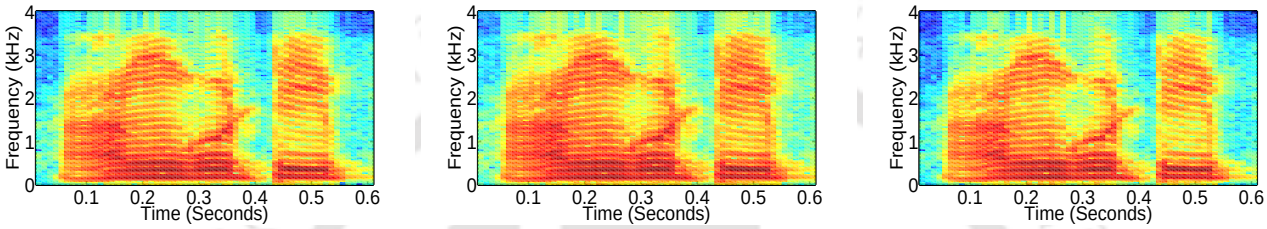
Speech recorded under happy condition: (m) original (n) global dictionary and (o) adaptive dictionary

Figure 5.5: The visual analysis of speech utterances of word /angoothi/ of SUSSC database by plotting their waveforms. The waveforms are plotted for the original raw speech and the synthesized speech determined using global dictionary and utterance-specific adaptive dictionary.



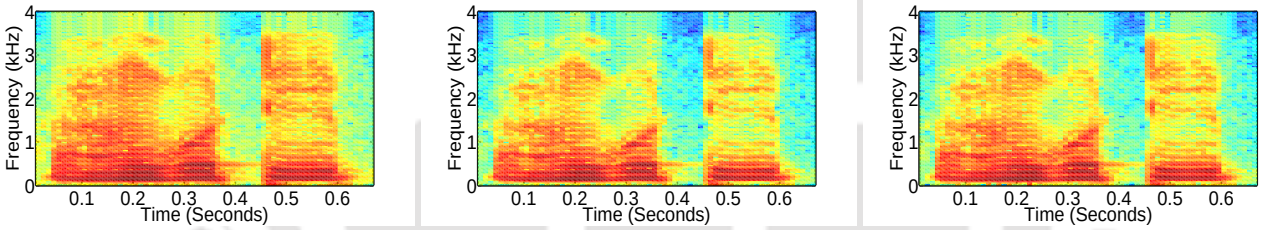
(a) (b) (c)

Speech recorded under neutral condition: (a) original (b) global dictionary and (c) adaptive dictionary



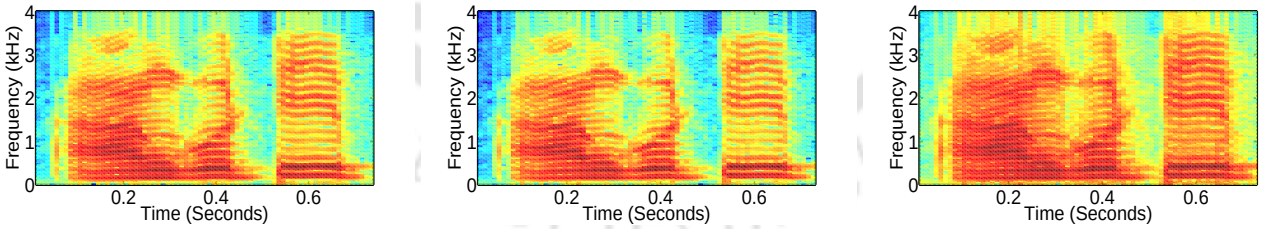
(d) (e) (f)

Speech recorded under angry condition: (d) original (e) global dictionary and (f) adaptive dictionary



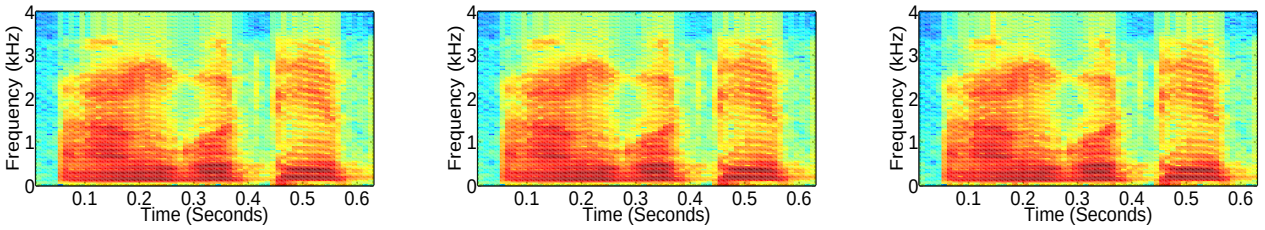
(g) (h) (i)

Speech recorded under sad condition: (g) original (h) global dictionary and (i) adaptive dictionary



(j) (k) (l)

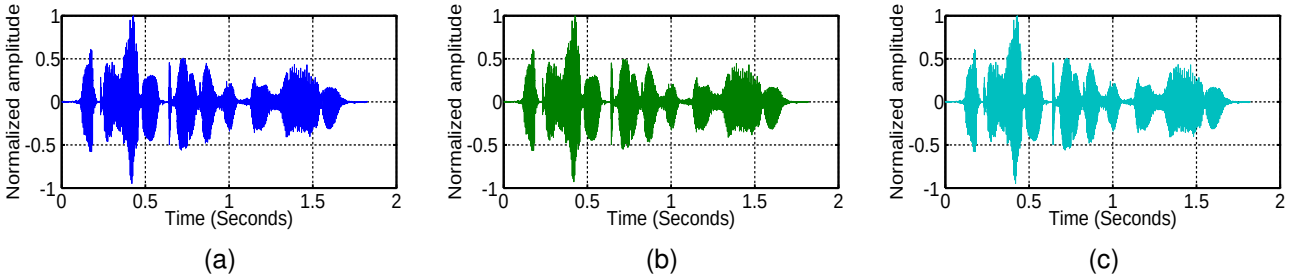
Speech recorded under lombard condition: (j) original (k) global dictionary and (l) adaptive dictionary



(m) (n) (o)

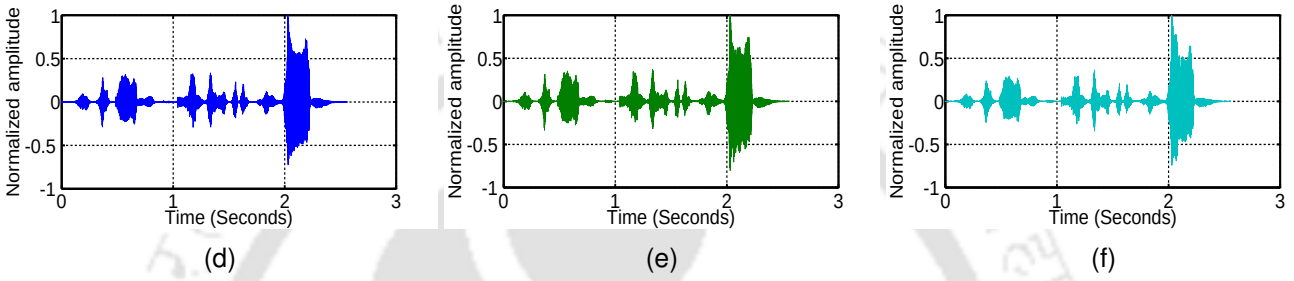
Speech recorded under happy condition: (m) original (n) global dictionary and (o) adaptive dictionary

Figure 5.6: The visual analysis of speech utterances of word /angoothi/ of SUSSC database by plotting their spectrograms. The spectrograms are plotted for the original raw speech and the synthesized speech determined using global dictionary and utterance-specific adaptive dictionary.



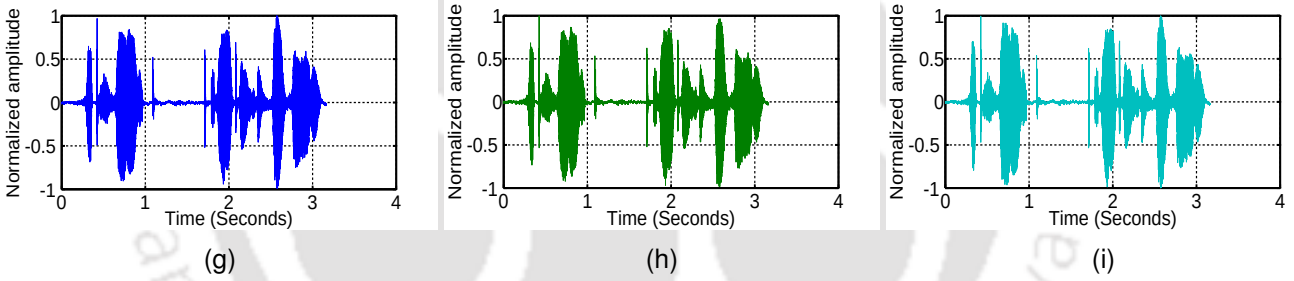
(a) (b) (c)

Speech recorded under neutral condition: (a) original (b) global dictionary and (c) adaptive dictionary



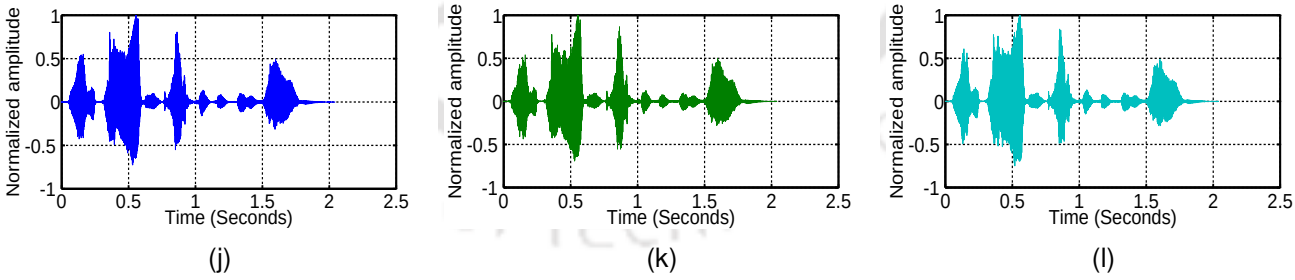
(d) (e) (f)

Speech recorded under angry condition: (d) original (e) global dictionary and (f) adaptive dictionary



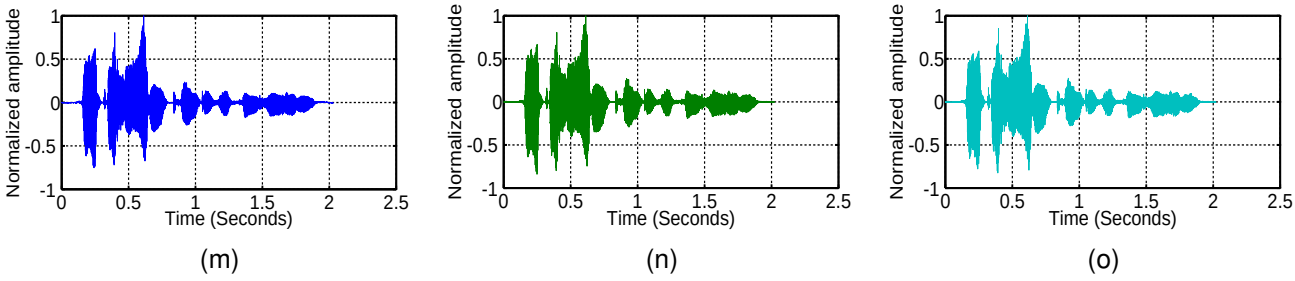
(g) (h) (i)

Speech recorded under sad condition: (g) original (h) global dictionary and (i) adaptive dictionary



(j) (k) (l)

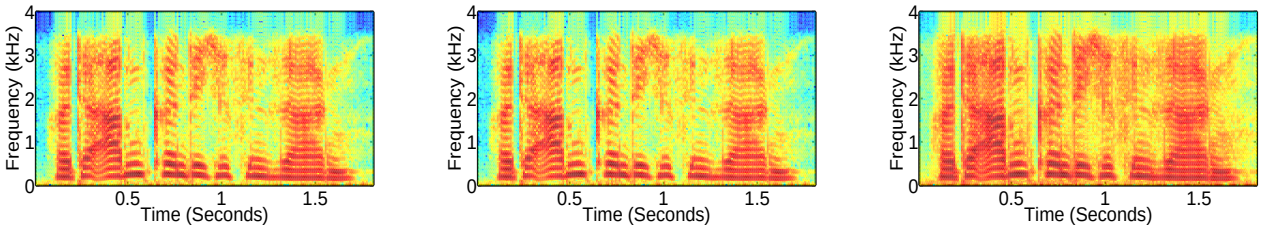
Speech recorded under happy condition: (j) original (k) global dictionary and (l) adaptive dictionary



(m) (n) (o)

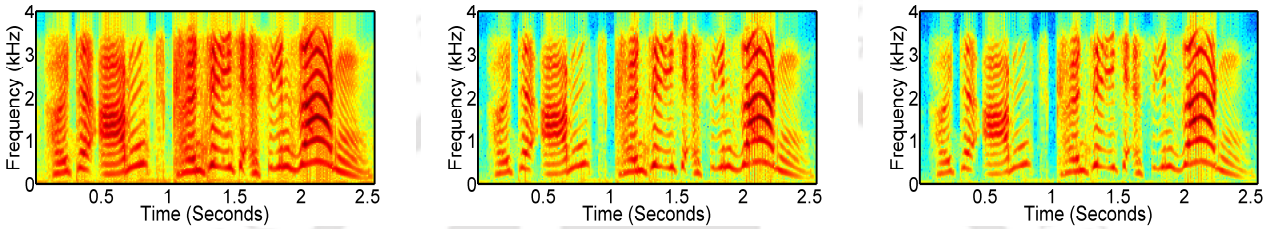
Speech recorded under boredom condition: (m) original (n) global dictionary and (o) adaptive dictionary

Figure 5.7: The visual analysis of speech utterances of sentence /Tonight I could tell him/ of Emo-DB database by plotting their waveforms. The waveforms are plotted for the original raw speech and the synthesized speech determined using global dictionary and utterance-specific adaptive dictionary.



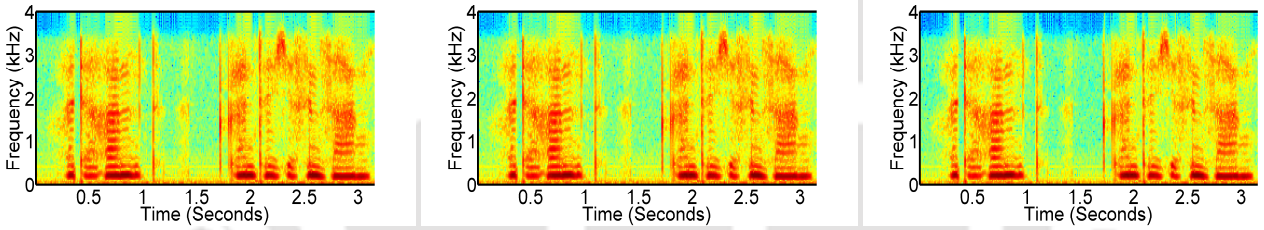
(a) (b) (c)

Speech recorded under neutral condition: (a) original (b) global dictionary and (c) adaptive dictionary



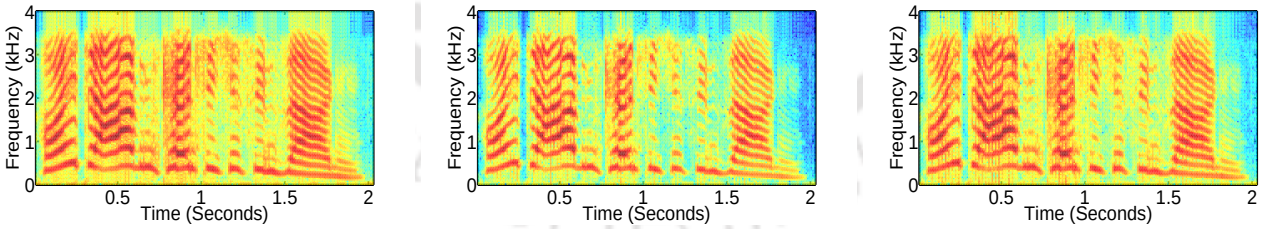
(d) (e) (f)

Speech recorded under angry condition: (d) original (e) global dictionary and (f) adaptive dictionary



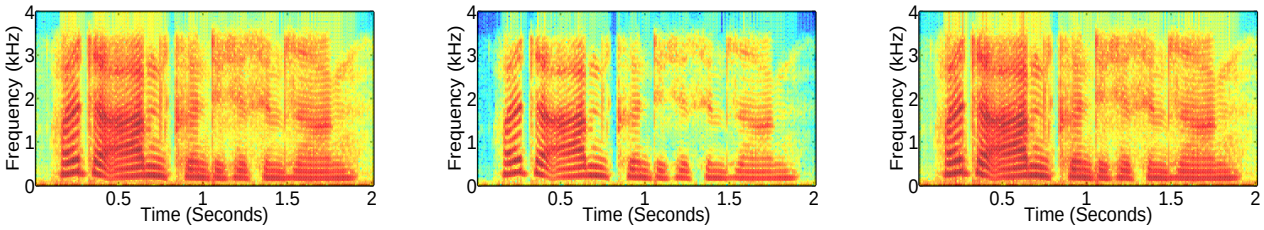
(g) (h) (i)

Speech recorded under sad condition: (g) original (h) global dictionary and (i) adaptive dictionary



(j) (k) (l)

Speech recorded under happy condition: (j) original (k) global dictionary and (l) adaptive dictionary



(m) (n) (o)

Speech recorded under boredom condition: (m) original (n) global dictionary and (o) adaptive dictionary

Figure 5.8: The visual analysis of speech utterances of sentence /Tonight I could tell him/ of Emo-DB database by plotting their spectrograms. The spectrograms are plotted for the original raw speech and the synthesized speech determined using global dictionary and utterance-specific adaptive dictionary.

measures the variations in the air pressure with respect to the time that human auditory system are able to perceive as sound. In frequency domain, the synthesized speech signals are quantified by studying the changes in their spectral distributions over the time. In this work, the variations in the air pressure and the spectral distribution are measured by interpreting the waveform and the spectrogram, respectively. Figure 5.5 and Figure 5.6 show the waveforms and the spectrograms for the synthesized speech utterances of word /angoothi/ of SUSSC database recorded from the same speaker under neutral, angry, sad, lombard and happy conditions, respectively. Using Emo-DB database, the waveforms and the spectrograms for the synthesized speech utterances of sentence /Tonight I could tell him/ produced from the same speaker under neutral, angry, sad, happy and boredom conditions are shown in Figure 5.7 and Figure 5.8, respectively. The synthesized speech signals are derived using the proposed stress normalization method by exploring both the invariable size global dictionary and the utterance-specific adaptive dictionary. In these figures, the waveforms and the spectrograms are also plotted for the original raw speech utterances in all the studied cases for comparison. It has been observed that, the proposed synthesis method using the global dictionary and the utterance-specific adaptive dictionary slightly changes the amplitude values of air pressure over the time. Consequently, the waveforms of synthesized speech utterances exhibit quite similar patterns with respect to the time in comparison to the waveforms of original raw speech utterances with little variation in the sound pressures. This trend is consistently observed for the neutral and the stressed speech utterances of both the studied databases as depicted in Figure 5.5 and Figure 5.7, respectively. The waveforms shown in Figure 5.5(g)−Figure 5.5(i) and Figure 5.5(m)−Figure 5.5(o) illustrate the changes in the air pressure approximately between 0.2 to 0.4 and 0.3 to 0.6 seconds in the original raw speech and the synthesized speech under sad and happy conditions of SUSSC database, respectively. Similarly using Emo-DB database, the maximum modifications in the air pressure are observed approximately between 2 to 3.5 and 0.5 to 3 seconds in the original raw utterances and the synthesized utterances of angry and sad speech as shown in Figure 5.7(d)−Figure 5.7(f) and Figure 5.7(g)−Figure 5.7(i), respectively. These experimental observations demonstrate that, the proposed synthesis technique preserve the original phonetic structures with little differences in the sound pressure and it helps in retaining the intelligibility of speech signals. The proposed sparse representation of vocal-tract system parameters for the neutral and the stressed speech over both the explored invariable size global dictionary and the utterance-specific adaptive dictionary diminish

the acoustic mismatch between them. Consequently, the synthesis of neutral and stressed speech by employing their corresponding estimated vocal-tract system parameters helps in comprising the similar acoustic characteristics as well as in maintaining their original patterns of sound over the time. The spectrograms of speech utterances of both the databases explored in this work illustrate the maximum changes between the spectral characteristics of original raw speech and synthesized speech in the moderate to high audio frequency bandwidth as depicted in Figure 5.6 and Figure 5.8, respectively. Using both the SUSSC and the Emo-DB databases, the maximum changes in the spectral distribution have been observed between 1 kHz to 2.5 kHz frequency bandwidth for the original raw angry speech and the synthesized angry speech as shown in Figure 5.6(d)−Figure 5.6(f) and Figure 5.8(d)−Figure 5.8(f), respectively. Whereas, the larger variations are appeared between 1.5 kHz to 3.5 kHz audio frequency range between the original sad speech and the synthesized sad speech over both the studied databases as depicted in Figure 5.6(g)−Figure 5.6(i) and Figure 5.8(i)−Figure 5.8(i), respectively. These variations in the spectral characteristics between the original raw and the synthesized speech utterances demonstrate that, the proposed synthesis process using both the proposed invariable size global dictionary and the utterance-specific adaptive dictionary suppress the divergences resulting from stress, which generally appear in the moderate to high frequency bandwidths and also preserves the original acoustic characteristics.

Error Analysis: The error analysis quantifies the synthesized speech signals by measuring the relative entropy between the Gaussian-subspaces by exploiting the Kullback Leibler (KL) divergence metric [140, 141]. The KL divergence values are determined using both the SUSSC and the Emo-DB databases and are summarized in the bar plots shown in Figure 5.9. The Gaussian-subspaces are developed using the GMMs comprising the mixture of 32 diagonal Gaussian densities. The experimental setup used to evaluate the relative entropies over both the databases explored in this work are similar to as described in the error analysis paragraph summarized in Section 4.2 in Chapter 4. The labeling of X-axis as 'Disg.' and 'Bore.' in the bar plot shown in Figure 5.9(b) denote the disgust and the boredom condition, respectively. In these bar plots, the relative entropy values in case of global dictionary and adaptive dictionary measure the KL divergences between the GMMs learned on the features of synthesized neutral and synthesized stressed speech utterances derived using the proposed stress normalization method employing the invariable size global dictionary and the utterance-specific adaptive dictionary, respectively. The relative entropy values summarized for the

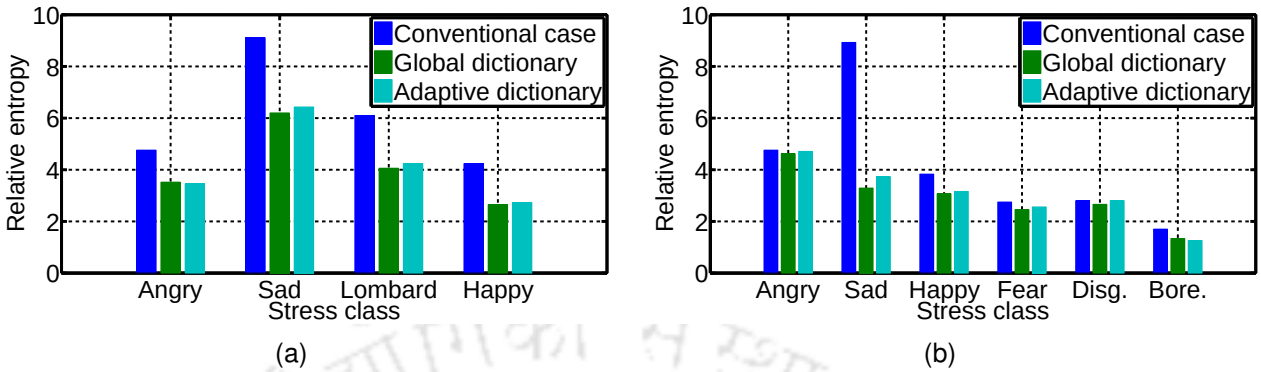


Figure 5.9: Evaluation of synthesized speech determined using the proposed stress normalization method employing global dictionary and utterance-specific adaptive dictionary by measuring the relative entropy between the Gaussain-subspaces using the KL divergence metric. (a) and (b) represent the KL divergence values with respect to the SUSSC and the Emo-DB databases, respectively.

conventional case are determined by measuring the KL divergence between the GMMs trained using the features of original raw utterances of neutral and stressed speech, respectively. The synthesis of speech utterances of angry, sad, lombard and happy speech of SUSSC database using the proposed synthesis process utilizing the global dictionary lead to the reduction in the KL divergence values by 26.05% (4.76 to 3.52), 31.94% (9.11 to 6.20), 33.44% (6.10 to 4.06) and 37.50% (4.24 to 2.65) in comparison to the KL divergences determined in conventional cases as shown in Figure 5.9(a). Whereas, Gaussian-subspaces developed on the synthesized angry, sad, lombard and happy speech utterances of SUSSC database determined using the proposed synthesis approach employing the utterance-specific adaptive dictionary has resulted in the relative decrement in the KL divergences of 27.31% (4.76 to 3.46), 29.42% (9.11 to 6.43), 30.49% (6.10 to 4.24) and 35.61% (4.24 to 2.73), when compared to the KL divergences obtained in the conventional cases, respectively. Similarly, using Emo-DB database, when speech signals are synthesized using the proposed stress normalization method employing the global dictionary, the KL divergence values for angry, sad, happy, fear, disgust and boredom speech are reduced by 2.94% (4.76 to 4.62), 63.16% (8.93 to 3.29), 19.84% (3.83 to 3.07), 10.54% (2.75 to 2.46), 5% (2.80 to 2.66) and 21.18% (1.70 to 1.34) in comparison to the KL divergences of conventional cases as shown in Figure 5.9(b). The GMMs trained using the synthesized angry, sad, happy, fear and boredom speech derived using the proposed synthesis method employing the utterance-specific adaptive dictionary yield the relative reduction in the KL divergences of 1.05% (4.76 to 4.71), 58.23% (8.93 to 3.73), 17.49% (3.83 to 3.16), 6.91% (2.75 to 2.56)

and 25.88% (1.70 to 1.26) in comparison to the KL divergence values obtained in the conventional cases, respectively. Moreover, the relative entropies reported in Figure 5.9 show that, using both the SUSSC and the Emo-DB databases, the proposed stress normalization method exploiting the invariable size global dictionary has reduced the relative entropies between the Gaussian-subspaces over all the studied stress classes by average value of 32.07% (6.05 to 4.11) and 29.54% (4.13 to 2.91) in comparison to the average values of KL divergences obtained in conventional cases, respectively. Whereas, the synthesis of speech signals of both the studied databases by employing the utterance-specific adaptive dictionary has resulted in the reduction in the relative entropies over all the explored stress classes by 30.41% (6.05 to 4.21) and 26.63% (4.13 to 3.03), when compared to the relative entropies determined in the conventional cases, respectively. These experimental results illustrate that, the proposed stress normalization method employing the sparse representation of vocal-tract system parameters of neutral and stressed speech over both the explored invariable size global dictionary and utterance-specific adaptive dictionary significantly reduces the acoustic mismatch between them. This, in turns, the synthesis of neutral and stressed speech using their corresponding estimated vocal-tract system parameters reduces the variance mismatch between the different speech units of neutral and stressed speech. Consequently, Gaussian subspaces developed using the synthesized neutral and the synthesized stressed speech exhibit the similar acoustic properties with the significantly reduced values of relative entropy between them.

5.3 Speaker Adaptation

This Section addresses the consequence of speaker variability for increasing the effectiveness of the proposed stress normalization method. In Chapter 4, the normalization of speaker-specific attributes from both the feature- and the model-space were found to be very effective for reducing the variance mismatch between the training and test environments of an automatic speech recognition system (ASR) system. The resulting speaker dependent (SD) ASR systems have been reported to be improved speech recognition performances for all the studied stress classes as depicted in Table 4.2-Table 4.5 and Figure 4.7−Figure 4.10. Motivated by those studies, in this work, once again, the feature- and the model-space are adapted onto another common subspace, which constitutes the speaker normalize information. The normalization of speaker variability can help in increasing the robustness of the proposed sparse representation of vocal-tract system parameters over the invariable

size global dictionary and the utterance-specific adaptive dictionary. The technique employed for the speaker adaptation is similar to as described in Section 4.3 and Section 4.4 in Chapter 4. For convenience purpose, a brief discussion on the used technique is presented here. At first, the synthesized speech features are spliced in time considering a context size of t_f. The dimensionality of the derived time-spliced features is then reduced to $L = 40$ by exploring the linear discriminant analysis (LDA)-based low-rank subspace projection [147] in the maximum likelihood linear transformation (MLLT) [148] technique. After that, the feature-space maximum likelihood linear regression (fMLLR) [144, 145, 161] transformations are generated for the training and the test speech utterances using the speaker adaptive training (SAT) [95, 97, 146, 162] approach. Furthermore, the dimension of feature- and model-space are varied below the default $L = 40$ dimension to measure the effectiveness of proposed stress normalization method onto the dimension lower than that of the default dimension.

5.4 Experimental Evaluation and Discussion

In this Section, the effectiveness of the proposed stress normalization technique exploring the sparse representation of vocal-tract system parameters with speaker normalization is measured using the speech recognition for the stressed speech. The performances of various ASR systems developed in this work are evaluated using the word error rate (WER) metric and is determined using Eq. 2.9 as given in Chapter 2. The following paragraphs describe the experimental setup used and the performance evaluation of the proposed stress normalization method.

Experimental Setup: The performance of the proposed stress normalization technique is quantified using the SUSSC database reported in [118] as summarized in Section 2.1 in Chapter 2. The speech parametrization technique for the speech utterances are similar to as described in the experimental setup Subsection 2.3.2 in Chapter 2. The LPCs are derived using the $P = 25$ order LP analysis. 2322 neutral speech utterances belonging to the training database are used for creating both the invariable size global dictionary and the utterance-specific adaptive dictionary for the sparse representation with a sparsity constraint of $T_0 = 10$. The global dictionary constitutes $D = 500$ atoms. The utterance-specific adaptive dictionary is constructed using $K_n = 20$ nearest neighbour LP-vectors with respect to the Euclidean distance metric. The number of atoms in the utterance-

specific adaptive dictionary vary as $D = K_n F = 20F$ with the F frames of speech utterance. In Chapter 4, the deteriorated performances of stressed speech recognition were reported over the SGMM-based acoustic modelling techniques as presented in Table 4.3, Table 4.5, Figure 4.10 and Figure 4.10. Therefore, in this study, in modelling paradigm, we have explored the speaker dependent ASR systems employing acoustic models based on Gaussian mixture model (GMM) as well as deep neural network (DNN). The remaining experimental setup is identical to as portrayed in Subsection 2.3.2 in Chapter 2.

Performance Evaluation: The speech recognition performances for stressed speech by exploring the proposed stress normalization method using the MFCC and the TEO-CB-Auto-Env features over the GMM-HMM and the DNN-HMM systems are tabulated in Table 5.1 and Table 5.2, respectively. The conventional case shown in these tables corresponds to the case of using baseline ASR system. The baseline ASR systems are developed using the 39-dimensional features of raw original neutral speech utterances. The WER values reported using the proposed stress normalization technique are determined corresponding to the default $L = 40$ dimensional features of the synthesized speech. These 40-dimensional features are computed from the time splicing of context size of 7, (t_f varies from $\pm$ 3) frames of the synthesized speech signals. The LDA+MLLT transformations are generated for the training and the test data to reduce the dimension of feature- and model-space upto 40 as well as to achieve the decorrelated characteristics. The fMLLR-based linear regression transformation is employed in SAT to normalize the speaker information. In these tables, the bold face WERs represent the best speech recognition performances in comparison to the conventional cases.

The performances of speech recognition have been reported to be degraded for the stressed speech cases in comparison to the neutral speech over both the explored feature and modelling paradigms as evident from the WERs summarized in Table 5.1 and Table 5.2, respectively. The proposed stress normalization technique using the K-SVD algorithm-based invariable size global dictionary and the K-NN algorithm-based utterance-specific adaptive dictionary are found to give competitive performances. It is to note that, the reported best stressed speech recognition performances corresponds to the utterance-specific adaptive dictionary. The proposed stress normalization method employing the utterance-specific adaptive dictionary using the GMM-HMM systems developed on the MFCC features has resulted in the maximum improvement in the recognition performance with the relative decrement in the WER values of 80.28%, 65.86% and 82.47%, when compared to the WERs

Table 5.1: The recognition performances for stressed speech (WER in %) using the proposed stress normalization technique employing sparse representation of LPCs with speaker adaptation using MFCC features with respect to the GMM-HMM and the DNN-HMM systems.

Modeling approach		Stress class	WER (in %)			% Relative reduction	
			Conventional	Proposed technique			
				Global dictionary	Adaptive dictionary	Global dictionary	Adaptive dictionary
GMM-HMM		Neutral	0.71	0.14	0.14	80.28	80.28
		Angry	25.71	7.57	9.86	**70.56**	61.65
		Sad	7.41	2.86	2.53	61.40	65.86
		Lombard	6.73	1.52	1.18	77.41	82.47
		Happy	7.86	2.14	3	72.77	61.83
DNN-HMM	1 hidden layer	Neutral	1.86	0.43	0.57	76.88	69.35
		Angry	22	8.14	9	63	59.09
		Sad	13.30	3.03	2.36	77.22	82.25
		Lombard	11.62	3.54	3.37	69.53	70.10
		Happy	11.14	3.86	4.14	65.35	62.84
	2 hidden layers	Neutral	1.43	0.14	0.14	90.21	90.21
		Angry	19	7.71	8.57	59.42	54.89
		Sad	11.28	2.19	80.58	80.58	80.58
		Lombard	9.26	1.68	2.53	81.86	72.68
		Happy	8.57	2.57	2.86	70.01	66.63
	3 hidden layers	Neutral	1.29	0.14	0.14	89.15	89.15
		Angry	18.86	7	7.86	62.88	58.32
		Sad	9.09	2.69	2.02	70.41	77.78
		Lombard	8.42	1.35	2.69	**83.97**	68.05
		Happy	8.57	2.71	2.71	68.38	68.38
	4 hidden layers	Neutral	1.29	0.14	0.14	89.15	89.15
		Angry	19.43	7.57	9.29	61.04	52.19
		Sad	10.44	2.36	2.36	77.39	77.39
		Lombard	7.41	1.85	2.53	75.03	65.86
		Happy	8.43	2.43	3	71.17	64.41
	5 hidden layers	Neutral	1.71	0.14	0.14	91.81	**91.81**
		Angry	18.86	7.43	8.57	60.60	54.56
		Sad	9.43	1.85	2.02	80.38	78.58
		Lombard	9.93	2.19	2.36	77.94	76.23
		Happy	8.43	2.71	2.29	67.85	72.83
	6 hidden layers	Neutral	1.57	0.14	0.29	91.08	81.53
		Angry	18.71	8.14	7.71	56.49	58.79
		Sad	9.93	1.85	1.35	81.37	**86.40**
		Lombard	9.43	1.85	2.36	80.38	74.97
		Happy	9	2.57	2.43	71.44	**73**

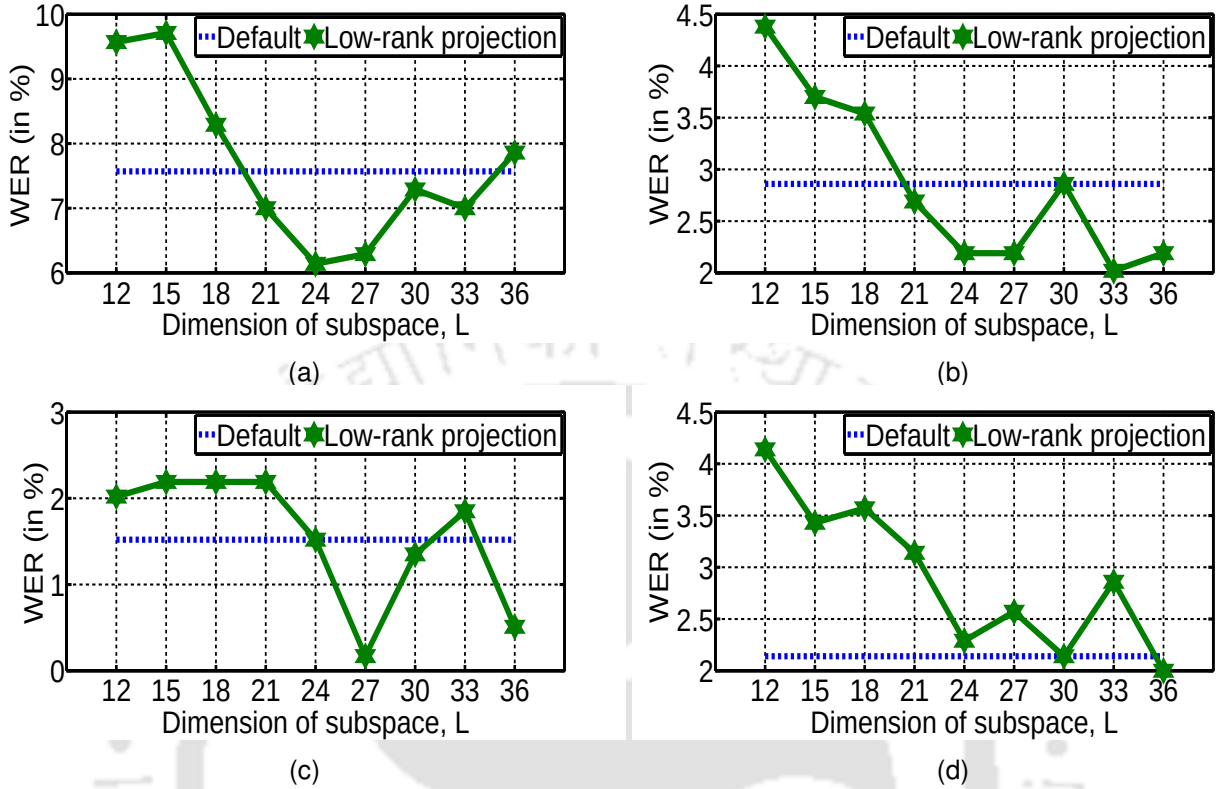
Table 5.2: The recognition performances for stressed speech (WER in %) using the proposed stress normalization technique employing sparse representation of LPCs with speaker adaptation using TEO-CB-Auto-Env features with respect to the GMM-HMM and the DNN-HMM systems.

Modeling approach		Stress class	WER (in %)			% Relative reduction	
			Conventional	Proposed technique			
				Global dictionary	Adaptive dictionary	Global dictionary	Adaptive dictionary
GMM-HMM		Neutral	0	0	0	**100**	**100**
		Angry	41.71	8.57	7.14	79.45	**82.88**
		Sad	31.31	6.57	7.58	79.02	75.79
		Lombard	24.07	3.87	3.70	83.92	**84.63**
		Happy	28.29	5.86	5.71	79.28	**79.82**
DNN-HMM	1 hidden layer	Neutral	1	0.29	0.14	71	86
		Angry	13.29	5.86	4.29	55.91	67.72
		Sad	14.31	5.05	4.21	64.71	70.58
		Lombard	7.91	2.19	2.02	72.31	74.46
		Happy	9.57	4.43	3.86	53.71	59.66
	2 hidden layers	Neutral	0.71	0.14	0.14	80.28	80.28
		Angry	12.14	4.57	4.57	62.35	62.35
		Sad	14.48	3.87	4.71	73.27	67.47
		Lombard	7.58	2.69	2.36	64.51	68.86
		Happy	8	4.43	3.14	44.62	60.75
	3 hidden layers	Neutral	0.71	0.14	0.14	80.28	80.28
		Angry	11.29	5.14	4.86	54.47	56.95
		Sad	12.96	4.88	5.05	62.34	61.03
		Lombard	7.41	2.36	1.52	68.15	79.49
		Happy	7.29	4.29	3.43	41.15	52.95
	4 hidden layers	Neutral	0.43	0.14	0.14	67.44	67.44
		Angry	12	4.86	5.57	59.5	53.58
		Sad	12.12	4.21	4.88	65.26	59.73
		Lombard	7.74	2.53	1.68	67.31	78.29
		Happy	7.57	3.57	3.71	52.84	50.99
	5 hidden layers	Neutral	0.57	0.14	0.14	75.44	75.44
		Angry	12.43	4.29	5.43	65.49	56.31
		Sad	11.95	4.55	5.05	61.92	57.74
		Lombard	7.91	3.03	2.36	61.69	70.16
		Happy	8	3.43	3.43	57.12	57.12
	6 hidden layers	Neutral	0.57	0.14	0.14	75.44	75.44
		Angry	12.71	5.29	5.71	58.38	55.07
		Sad	12.12	5.39	4.88	55.53	59.73
		Lombard	8.42	3.03	2.02	64.01	76.01
		Happy	8.29	3.14	3.14	62.12	62.12

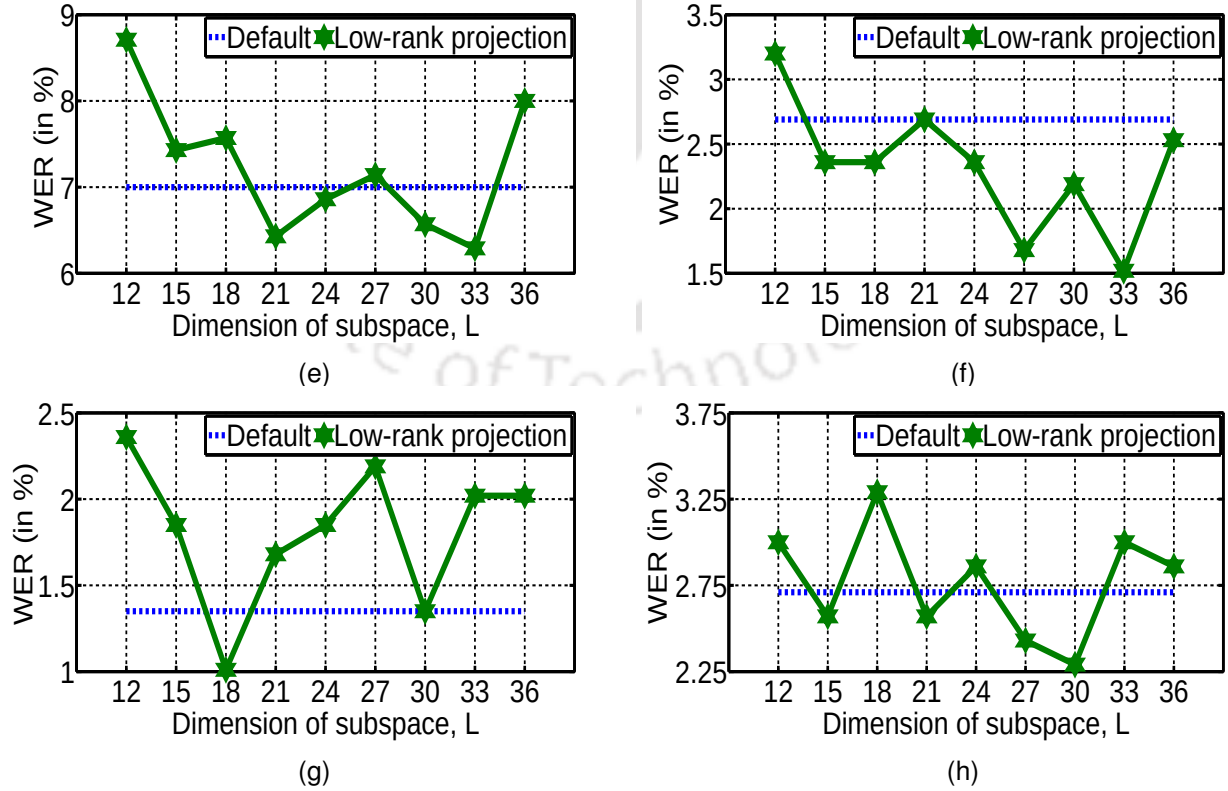
obtained in conventional cases for the recognition of neutral, sad and lombard speech, respectively, as depicted in Table 5.1. The similar effect is observed from the WER values presented in Table 5.2 for the TEO-CB-Auto-Env features. Using proposed stress normalization technique with utterance-specific adaptive dictionary, the maximum relative decrement in WERs of 100%, 82.88%, 84.63%, and 79.82% are obtained in comparison to the WER values determined using the baseline ASR systems, when GMM-HMM systems are developed on the TEO-CB-Auto-Env features for the recognition of neutral, angry, lombard and happy speech, respectively. Using DNN-based acoustic modelling technique, the performances of stressed speech recognition using the proposed stress normalization method over both the MFCC and the TEO-CB-Auto-Env features are evaluated by varying the number of hidden layers from 1 to 6 in DNN-HMM system as depicted in Table 5.1 and Table 5.2, respectively. Using utterance-specific adaptive dictionary, DNN-HMM systems comprising 5, 6 and 6 number of hidden layers, when trained on the MFCC features leaded to the best speech recognition performances for recognizing neutral, sad and happy speech, respectively. Using these specific number of hidden layers, the utterance-specific adaptive dictionary has resulted in the improvement in the recognition performance with the relative decrement in the WERs of 91.81%, 86.40% and 73% in comparison to the WER values obtained in conventional cases for the recognition of neutral, sad and happy speech, respectively. Similarly using TEO-CB-Auto-Env features, the utterance-specific adaptive dictionary yielded the best speech recognition performances with the relative decrement in the WER values of 86%, 67.72%, 79.49% and 62.12% over the DNN-HMM systems with 1, 1, 3 and 6 hidden layers, when compared to the WERs determined using the baseline ASR systems for recognizing the speech produced under neutral, angry, lombard and happy conditions, respectively. Except for some cases, the proposed stress normalization method employing the utterance-specific adaptive dictionary has established the improved speech recognition performances in comparison to the conventional cases and the case of using invariable size global dictionary over all the stress classes and the speech parameterizations approaches explored in this work. The reason of achieving improved performances despite such a large size of utterance-specific exemplar dictionary is attributed to the significant representation of the atoms, such that all the dimensions of the acoustic variations arises in LPCs are modeled using the K-NN algorithm-based probability density estimation method. Stress affects the muscles of vocal codes and the vocal tract and thus the proposed approach by incorporating the information about the duration parameter of speech signal provides an improved

dictionary creation method compared to the efficient implementation of well known K-SVD algorithm-based global dictionary. The sparse representation of LPCs for the neutral and the stressed speech over the utterance-specific adaptive dictionary reduces the acoustic mismatch between them. Consequently, the synthesis of neutral and stressed speech using their corresponding estimated LPCs helps in reducing the variance mismatch between them. Furthermore, the employed LDA-based low-rank subspace projection in MLLT-based semi-tied adaptation technique is appeared to be very effective for decorrelating the feature- and the model-space. The fMLLR-based adaptation of feature-space in SAT framework reduces the dissimilarities related to the speaker-specific information very effectively from the synthesized speech and the model paradigms. Moreover, the DNN-based ASR system with very small and large number of hidden layers has resulted in slightly degraded performances in comparison to that of the moderate number of hidden layers. These deteriorations are attributed to the elementary and the complex non-linear mapping between the input and the output of DNN-based acoustic modelling technique as discussed in Chapter 2 and Chapter 4.

As discussed in the above paragraph, the DNN-based ASR system with the adequate number of hidden layers effectively model the different speech units of synthesized neutral and synthesized stressed speech and it has resulted in convincing improvement in the recognition performances for the stressed speech. Therefore, in this study, the effectiveness of the proposed stress normalization technique by varying the dimension of feature- and model-space is investigated for the DNN-based ASR system for the 3 hidden layers. The performance of stressed speech recognition using the proposed method by exploring the low-rank subspace projection for all the studied acoustic features and modelling approaches are depicted by the WER-profiles shown in Figure 5.10$-$Figure 5.13. The WER-profiles obtained using the default case corresponds to the case with the 40-length features derived using the proposed method, as illustrated earlier. To obtain the WER-profiles by varying the dimension of feature- and model-space, the rank of the subspace projection matrix is varied from 36 to 12 in steps of 3. The performances of stressed speech recognition are measured by determining the WERs over a separate ASR system trained and tested with corresponding reduced dimensional features obtained from every reduction in the rank of the subspace projection matrix. The acronyms 'NI' shown in these tables correspond to the case of no relative improvement in the speech recognition performances in comparison to the default cases. The WER-profiles depicted in these figures demonstrate that, the subspace projection of both the synthesized neutral and the syn-

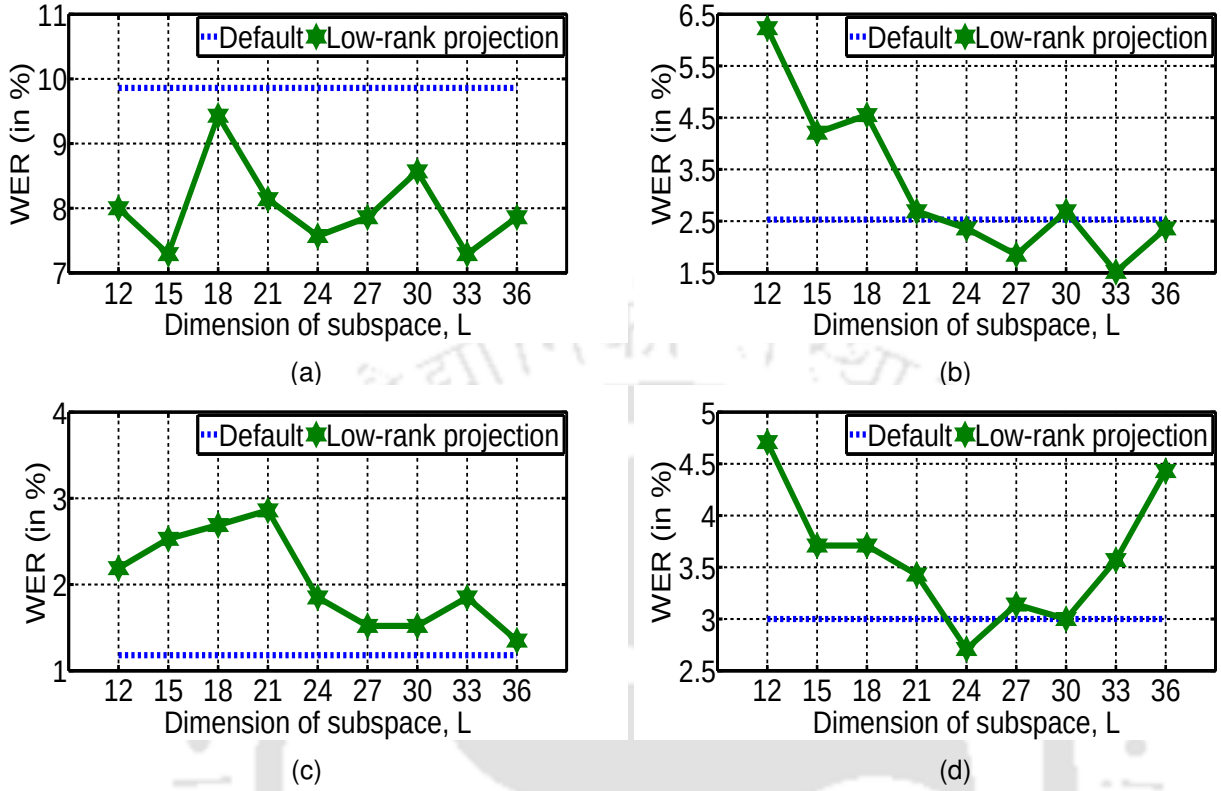


Low-rank projection over the GMM-HMM system; (a) angry, (b) sad, (c) lombard and (d) happy

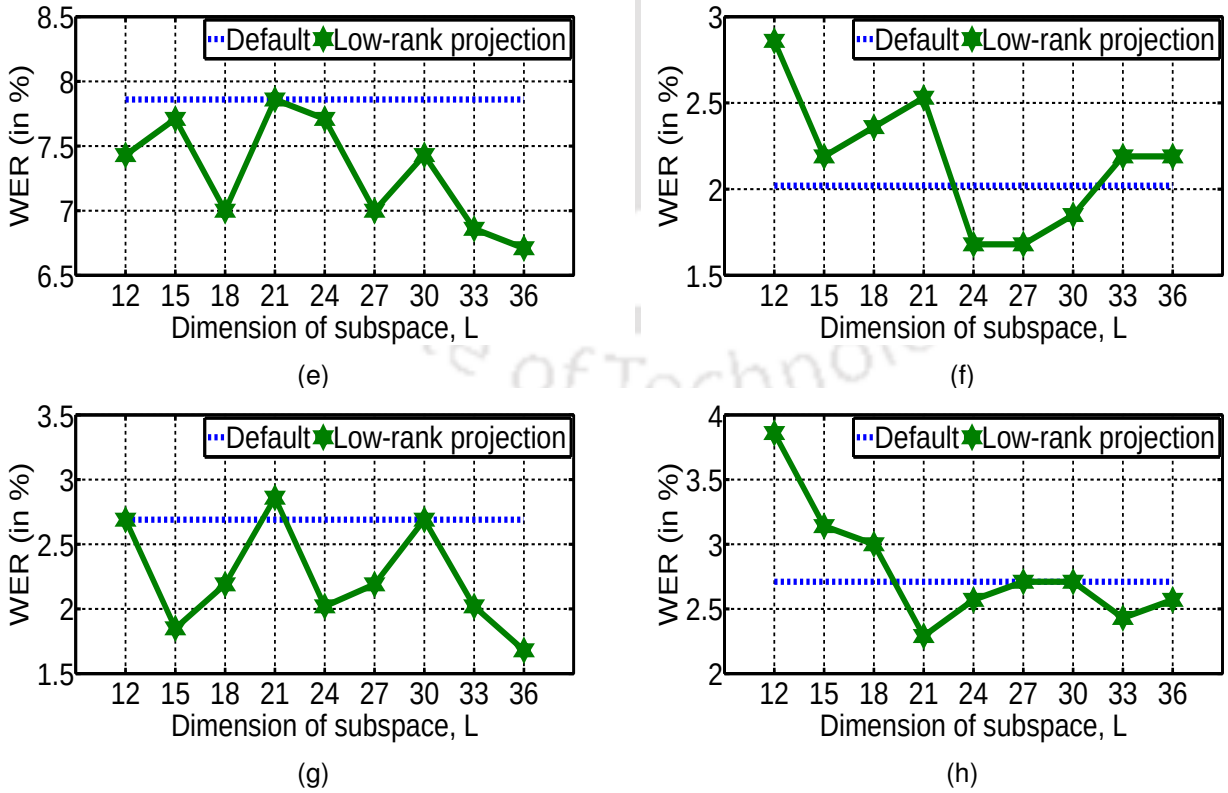


Low-rank projection over the DNN-HMM system; (e) angry, (f) sad, (g) lombard and (h) happy

Figure 5.10: Change in the WERs using the proposed sparse representation of LPCs employing the invariable size global dictionary with speaker adaptation with respect to the separate GMM-HMM and DNN-HMM systems developed on the MFCC features.

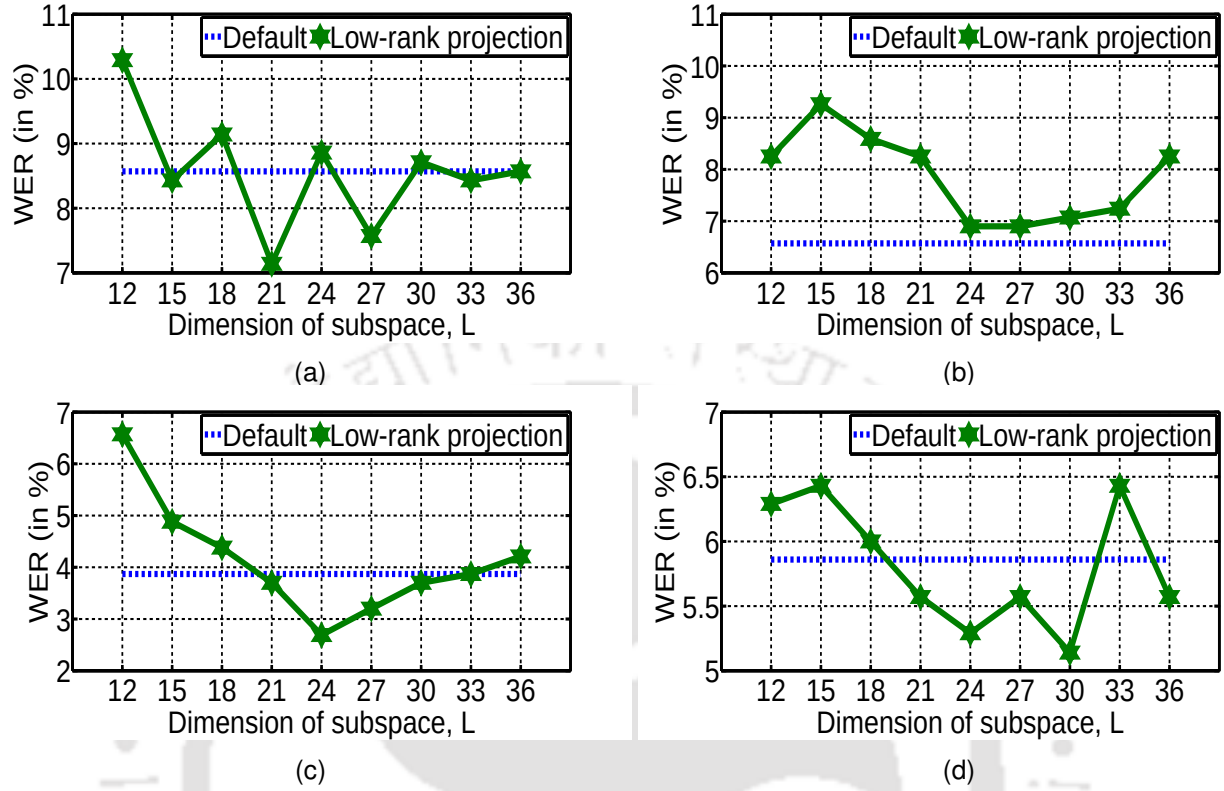


Low-rank projection over the GMM-HMM system; (a) angry, (b) sad, (c) lombard and (d) happy

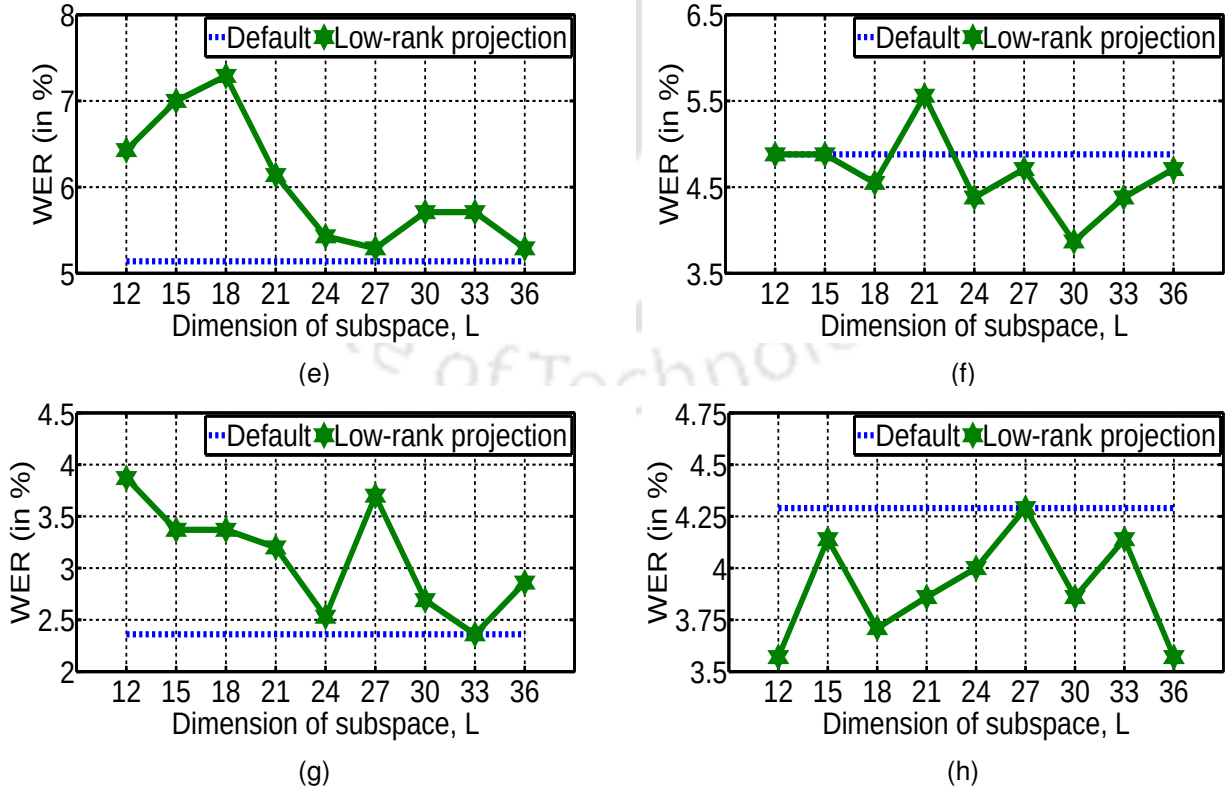


Low-rank projection over the DNN-HMM system; (e) angry, (f) sad, (g) lombard and (h) happy

Figure 5.11: Change in the WERs using the proposed sparse representation of LPCs employing the utterance-specific adaptive dictionary with speaker adaptation with respect to the separate GMM-HMM and DNN-HMM systems developed on the MFCC features.

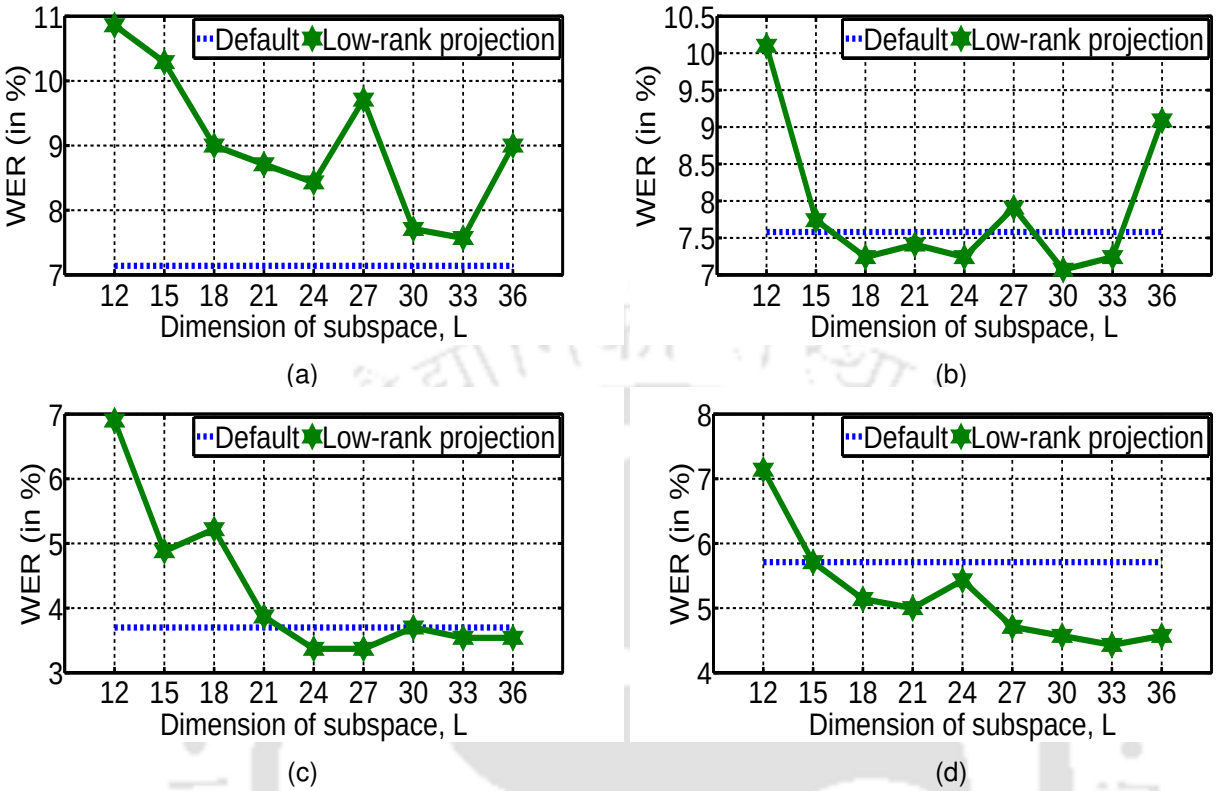


Low-rank projection over the GMM-HMM system; (a) angry, (b) sad, (c) lombard and (d) happy

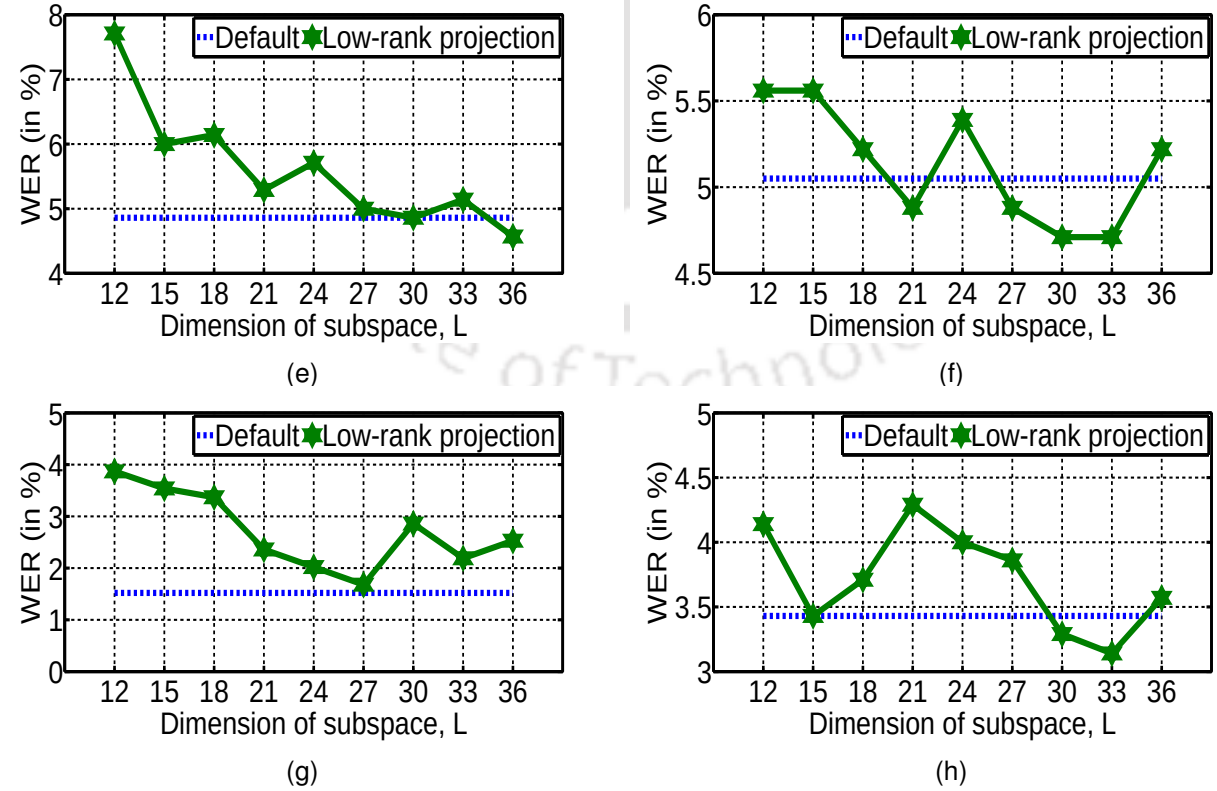


Low-rank projection over the DNN-HMM system; (e) angry, (f) sad, (g) lombard and (h) happy

Figure 5.12: Change in the WERs using the proposed sparse representation of LPCs employing the invariable size global dictionary with speaker adaptation with respect to the separate GMM-HMM and DNN-HMM systems developed on the TEO-CB-Auto-Env features.



Low-rank projection over the GMM-HMM system; (a) angry, (b) sad, (c) lombard and (d) happy



Low-rank projection over the DNN-HMM system; (e) angry, (f) sad, (g) lombard and (h) happy

Figure 5.13: Change in the WERs using the proposed sparse representation of LPCs employing the utterance-specific adaptive dictionary with speaker adaptation with respect to the separate GMM-HMM and DNN-HMM systems developed on the TEO-CB-Auto-Env features.

Table 5.3: Percentage of relative reductions in WERs obtained using the proposed sparse representation of LPCs with speaker adaptation technique. The performances are given for the MFCC features with respect to the GMM-HMM and the DNN-HMM systems. Default-rank subspace projection corresponds to the case of using 40-dimensional feature vector while best case performances are given for the low-rank subspace projection.

Modeling approach	Stress class	Global dictionary		Utterance-specific dictionary		% Relative reduction	
		Default-rank projection	Low-rank projection	Default-rank projection	Low-rank projection	Global dictionary	Utterance-specific dictionary
GMM-HMM	Angry	7.57	6.14	9.86	7.29	18.89	**26.06**
	Sad	2.86	2.02	2.53	1.52	29.37	39.92
	Lombard	1.52	0.17	1.18	1.35	**88.81**	NI
	Happy	2.14	2	3	2.71	6.54	9.67
DNN-HMM	Angry	7	6.29	7.86	6.71	10.14	14.63
	Sad	2.69	1.52	2.02	1.68	**43.49**	16.83
	Lombard	1.35	1.01	2.69	1.68	25.18	37.55
	Happy	2.71	2.29	2.71	2.29	15.50	**15.50**

thesized stressed speech signals onto a decorrelated feature subspace having dimension lower than that of the default $L = 40$ dimension, in which speaker information is also suppressed significantly diminishes the acoustic mismatch between them. Table 5.3 and Table 5.4 summarize the relative reduction in the WERs using the best performing low-rank features out of the default rank features for the MFFC and the TEO-CB-Auto-Env features with respect to the GMM-HMM and the DNN-HMM systems, respectively. On comparing the best performing $L = 15$ and $L = 27$ dimensional MFCC features out of the default-rank features, the proposed approach employing the utterance-specific dictionary and the global dictionary over the GMM-HMM system are noted to the maximum degradation of 26.06% and 88.81% in WERs for the recognition of angry and lombard speech as depicted in Figure 5.11(a) and Figure 5.10(c), respectively. Using proposed stress normalization method with global dictionary and utterance-specific dictionary, the DNN-HMM systems developed on the MFCC features having dimension $L = 33$ and $L = 21$ features yield the maximum relative improvement in the recognition performances by the decrement in the WERs of 43.49% and 15.50% for the recognition of sad and happy speech in comparison to the WER values obtained in default-rank subspace

Table 5.4: Percentage of relative reductions in WERs obtained using the proposed sparse representation of LPCs with speaker adaptation technique. The performances are given for the TEO-CB-Auto-Env features with respect to the GMM-HMM and the DNN-HMM systems. Default-rank subspace projection corresponds to the case of using 40-dimensional feature vector while best case performances are given for the low-rank subspace projection.

Modeling approach	Stress class	Global dictionary		Utterance-specific dictionary		% Relative reduction	
		Default-rank projection	Low-rank projection	Default-rank projection	Low-rank projection	Global dictionary	Utterance-specific dictionary
GMM-HMM	Angry	8.57	7.14	7.14	7.57	**16.69**	NI
	Sad	6.57	6.9	7.58	7.07	NI	6.73
	Lombard	3.87	2.69	3.70	3.37	**30.49**	8.92
	Happy	5.86	5.14	5.71	4.43	12.29	**22.42**
DNN-HMM	Angry	5.14	5.29	4.86	4.57	NI	5.97
	Sad	4.88	3.87	5.05	4.71	**20.70**	6.73
	Lombard	2.36	2.36	1.52	1.68	0	NI
	Happy	4.29	3.57	3.43	3.14	16.78	8.45

projection cases as shown in Figure 5.10(f) and Figure 5.11(h), respectively. Similarly, using GMM-HMM systems developed on the TEO-CB-Auto-Env features having dimension $L = 21$, $L = 24$ and $L = 33$ are reported to the best recognition performances with the relative decrement in the WERs of 16.69%, 30.49% and 22.42% for the recognition of angry, lombard and happy speech in comparison to the WERs obtained using the default-rank features as shown in Figure 5.12(a), Figure 5.12(c) and Figure 5.13(d), respectively. For the recognition of sad speech, proposed stress normalization method employing the global dictionary leads to the maximum relative improvement in the recognition performances with the decrement in the WER of 20.70% over the DNN-HMM system developed on the $L = 30$ dimensional TEO-CB-Auto-Env features, when compared to the default-rank subspace projection as shown in Figure 5.12(f). Once more, significant reductions in WERs are quite evident using the proposed stress normalization method employing the K-SVD algorithm-based invariable size global dictionary and the K-NN algorithm-based utterance-specific adaptive dictionary in the lower dimensional subspace in comparison to the WERs obtained using the default-rank subspace projection and the conventional cases except for some cases. These experimental observations il-

lustrate that, the proposed low-rank subspace projection method effectively discards the discrepancy induces due to the stress in the high-rank features. In low frequency bandwidth, the stressed speech carries the acoustic properties similar to as the acoustic properties of neutral speech. Therefore, the ASR system exhibits the similar characteristics between the training and test environments, when trained and tested using the low-rank decorrelated features of synthesized neutral and synthesized stressed speech with the reduced speaker variability, respectively. Consequently, resulting speaker dependent ASR systems have been reported in consistent improvement in the speech recognition performance in all the stress classes explored in this work with the significant reduction in the WER values, when compared to the WERs obtained using the baseline ASR systems. The proposed sparse representation of LPC with speaker normalization approach effectively reduces the variance mismatch between the spectral distribution of different speech units of neutral and stressed speech in the lower dimensional subspace.

5.5 Summary

Sparse representation of speech signals has been widely used for several decades, and consequently, the proposed stress normalization method described in this Chapter is well developed. The central focus in this Chapter has been the investigation on the modification in the vocal-tract system under stress condition in sparse domain for the development of an effective stress normalization technique. The parameters of vocal-tract system are modeled using the linear prediction coefficients (LPCs). The dissimilarities between the speech units of neutral and stressed speech are reduced using the synthesis process by exploring their corresponding estimated LPCs exhibiting the similar acoustic properties. To retrieve the similar acoustic properties for the LPCs of neutral and stressed speech, they are linearly transformed with sparsity over the dictionary, which consists of LPCs of neutral speech data. In general, the sparse representation models the speech signal using sparse linear combinations of atoms from a predetermined dictionary. Naturally, the designing of appropriate dictionary will result with better representation of LPCs, and lead eventually to robust stress normalization method. In the proposed approach, the sparse representation has been accomplished over the dictionaries created on the two distinct frameworks: the K-SVD algorithm-based invariable size global dictionary and the K-NN algorithm-based utterance-specific adaptive dictionary. The incorporation of information about the duration parameter of speech utterance by exploring the K-NN algorithm

has been noted to be very effective for the development of an effective dictionary referred to as the utterance-specific adaptive dictionary. These observations illustrate that, both the explored invariable size global dictionary and utterance-specific adaptive dictionary lead to the robust representation of LPCs for the neutral and the stressed speech signals. This, in turns, the synthesis of neutral and stressed speech using their corresponding estimated LPCs diminishes the variance mismatch between them. The ASR system developed on the synthesized neutral speech when tested using the synthesized stressed speech has been reported to show improved recognition performances with the significant deterioration in the WER values in comparison to the WERs determined in conventional cases for all stress classes explored in this work.

Furthermore, the difference in the variances of the speech units of synthesized neutral and synthesized stressed speech are reduced by projecting both of them onto another common decorrelated subspace. Experimental observations presented in this work demonstrate that, the adaptation of feature- and model-space onto the decorrelated subspace of comparatively lower dimension than that of the default dimension leads to the robust representation of speech units of synthesized speech signals. To accomplish the proposed subspace projection, at first, the features of synthesized speech are projected onto the decorrelated feature subspace of lower dimension using LDA-based low-rank subspace projection method. The MLLT-based semi-tied adaptation technique is used to further decorrelate the resulting features as well as to adapt the model-space onto the same decorrelated subspace. The proposed low-rank subspace projection matrix derived using the synthesized neutral speech utterances has been reported to be very effective in preserving the intelligibility of speech signals. This is followed by the fMLLR-based transformation in the SAT framework to reduce the speaker variabilities. The performances of speech recognition for the stressed speech presented in this work illustrate that, the subspace projection of the synthesized speech signals onto a decorrelated subspace having dimension lower than that of the default dimension, in which speaker information has been also suppressed leads to the development of an effective stress normalization method.

6

Conclusions

Contents

This thesis proposes new schemes for stress normalization using the speech subspace modelling with speaker adaptation technique. The proposed stress normalization methods are the integration of three major endeavors namely: (i) development of effective speech subspace, (ii) normalization of speaker variability and (iii) low-rank subspace projection. Excellent progress has been achieved in studying the stressed speech, the reliability and the performance of stress normalization far short of human recognition and perception system. The normalization of stress information has been widely addressed by many speech research communities.

The Speaker changes the speech production system to acknowledge the information about the adverse environmental factors as well as to retain the intelligibility of speech signal. The speech produced under stress condition, which is generally referred to as the stressed speech exhibits the modified acoustic properties compared to the speech produced under neutral or normal condition, called as the neutral speech. Stress induces a high overlap between the different speech units of speech signals and it creates a large variance mismatch between the neutral and the stressed speech. Consequently, many contemporary, large scale laboratory and commercial applications, which involve their interaction with machines exhibit highly degraded performances for the users under stress conditions. For robust and reliable operation of these practical applications, it is required to have normalize the acoustic mismatch between the neutral and the stressed speech. Although several signal processing and pattern matching algorithms with promising results are proposed to analyze the stressed speech, there are still a number of fundamental problems which have motivated this work. The technique should be of sufficient quality to ensure the robust and the precise representation of both the neutral and the stressed speech. Among various techniques, subspace projection based methods have received a great deal of attention over the past several years. This thesis has mainly focussed on the development of subspace projection-based approaches for normalizing the stress information. The following Sections address the outline of important findings and the major contributions drawn from the works proposed in this thesis.

6.1 Outline of Important Findings

In this dissertation, three distinct approaches are proposed for reducing the acoustic mismatch between the neutral and the stressed speech. Based on the interpretations described in the previous Chapters, the important findings of works presented in this thesis can be outlined as follows.

Chapter 1 introduced the stressed speech and the significance of stress normalization. The literature survey on the processing of stressed speech in the acoustic-, the feature- and the model-space from existing published literatures are presented. The study and analysis of different subspace modelling techniques for the investigation on the changes in the properties of stressed speech are also described. In this Chapter, we have also discussed the various approaches proposed in literature to address the consequence of speaker variability in the production of stressed speech.

In **Chapter 2**, a brief discussion on the database used for the investigation on the characteristics of stressed speech is presented. The literature reviews on the extraction of features and the optimization of fine tuned model parameters for the robust and the precise representation of stressed speech onto the feature- and the model-space, respectively, are discussed. In this Chapter, the studied speech features and acoustic modelling techniques are evaluated for the stressed speech. Though many algorithms and proposition have been reported in the literature, it was observed that, they dedicated on studying the changes in the speech production system under stress condition. The investigation on the acoustic mismatch between the neutral and the stressed speech from the point of view of development of speech subspace and the speaker variability were not addressed. This gave us the motivation for designing a robust and computationally efficient stress normalization method by exploring the subspace projection of neutral and stressed speech onto a common subspace consisting of speech information with suppressed speaker variabilities.

Chapter 3 proposed the method based on the linear and the non-linear subspace modelling approaches for the development of effective stress normalization techniques. It utilizes the projection of neutral and stressed speech onto a common subspace, which consists of properties of neutral speech data. The proposed linear subspace has been developed by exploring the orthogonal projection and the linear transformation techniques. This study investigates the non-linearity between the speech and the stress information onto the non-linear data space by accomplishing the subspace projection through the non-linear transformation using the polynomial function. The effectiveness of proposed stress normalization methods are validated by varying the dimension of an effective speech subspace by exploring the heteroscedastic linear discriminant analysis (HLDA)-based low-rank subspace projection in the maximum likelihood linear transformation (MLLT)-based semi-tied adaptation framework. The main conclusions of the work, in this chapter, are as follows.

- The hypothesis of linearity between the speech- and the stress-specific attributes has been

found very fascinating for the filtering of an effective speech subspace.

- Orthogonal projection of stressed speech features onto the specific number of basis vectors of filtered speech subspace reduces the acoustic mismatch between the neutral and the stressed speech signals.
- The polynomial function of specific order has been shown effective for studying the non-linear relationship between the speech and the stress information.
- The non-linear subspace modelling using the polynomial function of specific order is appeared very effective for normalizing the stress information compared to the linear subspace modelling techniques derived using the orthogonal projection and the linear transformation approaches.
- The low-rank projection helps in discarding the variance mismatch resulting from stress, which generally appears in the high audio frequency bandwidths. It has been noted that, the stressed speech exhibits the properties similar to the neutral speech in the low frequency region.

In **Chapter 4**, the vocal-tract system is studied in the Gaussian-subspace for the normalization of stress-specific attributes. The vocal-tract system parameters of neutral and stressed speech are projected onto a common Gaussian-subspace, whose bases span the vocal-tract system parameters of neutral speech signals. In this work, we have derived the subspace projection by introducing the posterior probability information and it has estimated the posteriorgram features. These posteriorgram features are reported to be very effective in reducing the variance mismatch between the vocal-tract system parameters of neutral and stressed speech signals. Both the neutral and the stressed speech are synthesized using their corresponding estimated vocal-tract system parameters by exploring the LP coding technique. The synthesized speech are considered as the speech with the characteristics similar to the neutral speech. In this Chapter, the speaker normalization is also addressed to increase the effectiveness of the proposed stress normalization method. The feature-space maximum-likelihood linear regression (fMLLR)-based transformation is employed in the speaker adaptive training (SAT) mode to reduce the speaker variabilities from the synthesized speech. Furthermore, the dimension of feature- and model-space are varied below the default $L = 40$ dimension using the linear discriminant analysis (LDA)-based low-rank subspace projection in the MLLT-based semi-tied adaptation mode to measure the effectiveness of proposed stress normalization method onto the lower dimensional subspace. The main conclusions of this proposed work are as follows.

- Our study shows that, the development of Gaussian-subspace using the vocal-tract system parameters of neutral speech utterances estimates the vocal-tract system parameters for the stressed speech with the characteristics similar to the neutral speech.
- The synthesis of neutral and stressed speech using their corresponding estimated posteriorgram features with respect to their vocal-tract system parameters is reported to be very effective in reducing the variance mismatch between them.
- The experimental observations presented demonstrate that, the subspace projection of the synthesized speech signals onto a lower dimensional decorrelated subspace, in which speaker information has been also suppressed leads to the significant improvements in the speech recognition performances for the stressed speech.

Chapter 5 investigated the modification in the characteristics of vocal-tract system under stress condition in the sparse domain. A novel subspace modelling technique by introducing the sparse representation of vocal-tract system parameters for the neutral and the stressed speech is explored over the dictionary constituting the vocal-tract system parameters of neutral speech utterances. In this work, the sparse representation has been accomplished over the dictionaries created using the two distinct learning frameworks namely: (1) the K-SVD algorithm-based invariable size global dictionary and (2) the K-nearest-neighbour (K-NN) algorithm-based utterance-specific adaptive dictionary. The first learning mechanism exploits the jointly update of dictionary atoms along with update of sparse coding using the well known K-SVD algorithm for the creation of invariable size global dictionary. The utterance-specific adaptive dictionary is estimated by incorporating the information about the duration parameter of speech utterance by exploring the K-NN algorithm-based non-parametric probability density estimation method. The synthesis of neutral and stressed speech signals by employing their corresponding estimated vocal-tract system parameters results in speech signals constituting the similar acoustic properties. Furthermore, the LDA-based low-rank subspace projection is employed in the MLLT-based semi-tied adaptation technique to adapt the feature- and model-space onto a common decorrelated subspace having dimension lower than that of the default dimension. The speaker variability is also normalized by generating the fMLLR transformations for the training and test speech utterances using the SAT approach. The main conclusions are provided below.

- The observations from our experiments illustrate that, both the explored invariable size global

dictionary and utterance-specific adaptive dictionary have resulted in the robust representation of vocal-tract system parameters for the neutral and the stressed speech signals.

- The sparse representation over the utterance-specific adaptive dictionary has been reported to be improved stressed speech recognition performances compared to the conventional cases and the case of using invariable size global dictionary. The utterance-specific adaptive dictionary significantly models all the dimensions of acoustic variations arises in vocal-tract system parameters by incorporating the information about the duration parameter of speech signal.
- The speaker normalization has been reported to be very effective in increasing the robustness of the proposed stress normalization method. It has been observed that, in low frequency region, stressed speech exhibits the acoustic properties similar to as the neutral speech.

6.2 Major Contributions of the Work

The important contributions of the work reported in this thesis are as follows:

(i) The modelling of linear subspace by exploring the orthogonal projection-based subspace projection technique to investigate the linear relationship between the speech and the stress information for reducing the acoustic mismatch between the neutral and the stressed speech.

(ii) Development of effective stress normalization technique by investigating the non-linearity between the speech- and the stress-specific attributes onto the non-linear data space created by exploiting the polynomial function of specific order.

(iii) Normalization of stress information by introducing the posteriorgram representation of vocal-tract system parameters by exploring the subspace projection onto a common Gaussian-subspace, which incorporates the posterior probability information.

(iv) Designing of stress normalization algorithm by investigating the changes in the vocal-tract system in sparse domain using the sparse representation of vocal-tract system parameters over both the proposed invariable size global dictionary and utterance-specific adaptive dictionary.

The other contributions are,

(i) Speaker normalization is employed to reduce the acoustic mismatch resulting from the different speakers exhibiting different anatomical and physiological states under stress conditions.

(ii) The low-rank subspace projection has been explored to investigate the effect of stress in different audio frequency bands.

(iii) In modelling paradigm, ASR systems by exploring acoustic modelling techniques based on GMM, SGMM and DNN have been developed for the evaluation of proposed stress normalization methods. Additionally, the waveform, the spectral distribution, the KL divergence have been explored to measure the acoustic mismatch between the neutral and the stressed speech.

6.3 Scope for Future Research

From the result of the investigations carried out in this dissertation, the following topics have been found out for future research in this area:

(i) The proposed speech subspace modelling with speaker adaptation has been effective towards the presence of stress. The effect of pathological condition on the production of speech can also be studied and accordingly a more robust stress normalization technique can be designed.

(ii) The present thesis has primarily focussed on the designing of stress normalization methods by studying the changes in the vocal-tract system. The work may also be expanded to develop effective techniques for normalizing the stress information by analyzing the excitation source.

(iii) The proposed posteriorgram representation and sparse modelling based methods have used the subspace modelling method, which derives the subspace projection with respect to the full-rank and the over-complete subspace projection matrix, respectively. An effective stress normalization method can be developed which accounts for joint multi-view subspace projection matrix for each stress class to normalize the stress information.

(iv) In this dissertation, all the stress normalization methods are reported very effective in case of availability of large training and test data. The work may also be extended to design the stress normalization algorithm in the low-data scenario.

(v) The speaker normalization is reported to be very effective in increasing the robustness of stress normalization methods proposed in this thesis. The prosodic features of speakers can be introduced to model the more precisely the speaker identity for the development of an effective stress normalization technique.

References

[1] L. Rabiner and B.-H. Juang, *Fundamentals of speech recognition*. Prentice hall, 1993.

[2] J. G. Proakis and D. G. Manolakis, *Introduction to digital signal processing*. Prentice Hall Professional Technical Reference, 1988.

[3] L. R. Rabiner and R. W. Schafer, *Digital processing of speech signals*. IET, 1979, vol. 19.

[4] S. E. Bou-Ghazale and J. H. Hansen, "Generating stressed speech from neutral speech using a modified celp vocoder," *Speech Communication*, vol. 20, no. 1-2, pp. 93–110, 1996.

[5] C. E. Williams and K. N. Stevens, "Emotions and speech: Some acoustical correlates," *The Journal of the Acoustical Society of America*, vol. 52, no. 4B, pp. 1238–1250, 1972.

[6] J. H. Hansen and S. Patil, "Speech under stress: Analysis, modeling and recognition," in *Speaker Classification I*. Springer, 2007, pp. 108–137.

[7] H. Steeneken and J. Hansen, "Speech under stress conditions: overview of the effect on speech production and on system performance," in *Acoustics, Speech, and Signal Processing, 1999. Proceedings., 1999 IEEE International Conference on*, vol. 4, 1999, pp. 2079–2082 vol.4.

[8] W. H. Organization *et al.*, "Development of a global mental health action plan 2013–2020," 2014.

[9] I. R. Murray, J. L. Arnott, and E. A. Rohwer, "Emotional stress in synthetic speech: Progress and future directions," *Speech Communication*, vol. 20, no. 1-2, pp. 85–91, 1996.

[10] G. Fant and J. Lindquist, "Pressure and gas mixture effects on diver's speech," *Quart. Prog. and Status Rep. STL-QPSR, Speech Transmission Laboratory, Royal Institute of Technology, Stockholm, Sweden*, pp. 7–17, 1968.

[11] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition*, vol. 44, no. 3, pp. 572–587, 2011.

[12] M. Lew, E. M. Bakker, N. Sebe, and T. S. Huang, "Human-computer intelligent interaction: a survey," in *International Workshop on Human-Computer Interaction*. Springer, 2007, pp. 1–5.

[13] J. H. Hansen, "Analysis and compensation of stressed and noisy speech with application to robust automatic recognition," *Signal Processing*, vol. 17, no. 3, p. 282, 1989.

[14] R. W. Picard and R. Picard, *Affective computing*. MIT press Cambridge, 1997, vol. 252.

[15] J. Tao and T. Tan, "Affective computing: A review," in *International Conference on Affective computing and intelligent interaction*. Springer, 2005, pp. 981–995.

[16] B. Schuller, A. Batliner, S. Steidl, and D. Seppi, "Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge," *Speech Communication*, vol. 53, no. 9, pp. 1062–1087, 2011.

[17] H. Lu, D. Frauendorfer, M. Rabbi, M. S. Mast, G. T. Chittaranjan, A. T. Campbell, D. Gatica-Perez, and T. Choudhury, "Stresssense: Detecting stress in unconstrained acoustic environments using smartphones," in *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*. ACM, 2012, pp. 351–360.

[18] P. G. Hunter, E. G. Schellenberg, and A. T. Griffith, "Misery loves company: mood-congruent emotional responding to music." *Emotion*, vol. 11, no. 5, p. 1068, 2011.

[19] S. Bou-Ghazale and J. Hansen, "A comparative study of traditional and newly proposed features for recognition of speech under stress," *Speech and Audio Processing, IEEE Transactions on*, vol. 8, no. 4, pp. 429–442, Jul 2000.

[20] K. Wang, N. An, B. N. Li, Y. Zhang, and L. Li, "Speech emotion recognition using fourier parameters," *Affective Computing, IEEE Transactions on*, vol. 6, no. 1, pp. 69–75, Jan 2015.

[21] S. Shukla, S. Dandapat, and S. Prasanna, "Spectral slope based analysis and classification of stressed speech," *International Journal of Speech Technology*, vol. 14, no. 3, p. 245, 2011.

[22] T. Nwe, S. W. Foo, and C. De Silva, "Detection of stress and emotion in speech using traditional and fft based log energy features," in *Proceedings of the 2003 Joint Conference of the Fourth International Conference on*, vol. 3, 2003, pp. 1619–1623 vol.3.

[23] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor, "Emotion recognition in human-computer interaction," *IEEE Signal Processing Magazine*, vol. 18, no. 1, pp. 32–80, Jan 2001.

[24] E. H. Kim, K. H. Hyun, S. H. Kim, and Y. K. Kwak, "Improved emotion recognition with a novel speaker-independent feature," *IEEE/ASME Transactions on Mechatronics*, vol. 14, no. 3, pp. 317–325, June 2009.

[25] X. Yao, T. Jitsuhiro, C. Miyajima, N. Kitaoka, and K. Takeda, "Modeling of physical characteristics of speech under stress," *IEEE Signal Processing Letters*, vol. 22, no. 10, pp. 1801–1805, Oct 2015.

[26] J. Sundberg, S. Patel, E. Bjorkner, and K. R. Scherer, "Interdependencies among voice source parameters in emotional speech," *IEEE Transactions on Affective Computing*, vol. 2, no. 3, pp. 162–174, July 2011.

[27] B. Womack and J. Hansen, "N-channel hidden markov models for combined stressed speech classification and recognition," *Speech and Audio Processing, IEEE Transactions on*, vol. 7, no. 6, pp. 668–677, Nov 1999.

[28] J. C. Lin, C. H. Wu, and W. L. Wei, "Error weighted semi-coupled hidden markov model for audio-visual emotion recognition," *IEEE Transactions on Multimedia*, vol. 14, no. 1, pp. 142–156, Feb 2012.

[29] M. Afify, Y. Gong, and J. P. Haton, "A general joint additive and convolutive bias compensation approach applied to noisy lombard speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 6, pp. 524–538, Nov 1998.

[30] E. Vyrynen, J. Kortelainen, and T. Seppnen, "Classifier-based learning of nonlinear feature manifold for visualization of emotional speech prosody," *IEEE Transactions on Affective Computing*, vol. 4, no. 1, pp. 47–56, Jan 2013.

[31] Z. Zhang, E. Coutinho, J. Deng, and B. Schuller, "Cooperative learning and its application to emotion recognition from speech," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 115–126, Jan 2015.

[32] Y. Zong, W. Zheng, T. Zhang, and X. Huang, "Cross-corpus speech emotion recognition based on domain-adaptive least-squares regression," *IEEE Signal Processing Letters*, vol. 23, no. 5, pp. 585–589, May 2016.

[33] S. Yun and C. D. Yoo, "Loss-scaled large-margin gaussian mixture models for speech emotion classification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 2, pp. 585–598, Feb 2012.

[34] E. A. P. Habets and J. Benesty, "Multi-microphone noise reduction based on orthogonal noise signal decompositions," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 6, pp. 1123–1133, June 2013.

[35] K. Ozeki and T. Umeda, "An adaptive filtering algorithm using an orthogonal projection to an affine subspace and its properties," *Electronics and Communications in Japan (Part I: Communications)*, vol. 67, no. 5, pp. 19–27, 1984. [Online]. Available: http://dx.doi.org/10.1002/ecja.4400670503

[36] S. G. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Transactions on Signal Processing*, vol. 41, no. 12, pp. 3397–3415, Dec 1993.

[37] G. M. Davis, S. G. Mallat, and Z. Zhang, "Adaptive time-frequency decompositions," *Optical Engineering*, vol. 33, no. 7, pp. 2183–2191, 1994.

[38] B. Huang, S. X. Ding, and S. J. Qin, "Closed-loop subspace identification: an orthogonal projection approach," *Journal of process control*, vol. 15, no. 1, pp. 53–66, 2005.

[39] J. C. Harsanyi and C. I. Chang, "Hyperspectral image classification and dimensionality reduction: an orthogonal subspace projection approach," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 32, no. 4, pp. 779–785, Jul 1994.

[40] S. Shukla, S. Dandapat, and S. Prasanna, "Subspace projection based analysis of speech under stressed condition," in *Information and Communication Technologies (WICT), 2012 World Congress on*, Oct 2012, pp. 831–834.

[41] S. Ramamohan and S. Dandapat, "Sinusoidal model-based analysis and classification of stressed speech," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 3, pp. 737–746, May 2006.

[42] J.-C. Wang, Y.-H. Chin, B.-W. Chen, C.-H. Lin, and C.-H. Wu, "Speech emotion verification using emotion variance modeling and discriminant scale-frequency maps," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 23, no. 10, pp. 1552–1562, Oct 2015.

[43] L. Zao, D. Cavalcante, and R. Coelho, "Time-frequency feature and ams-gmm mask for acoustic emotion classification," *Signal Processing Letters, IEEE*, vol. 21, no. 5, pp. 620–624, May 2014.

[44] G. Zhou, J. H. Hansen, and J. F. Kaiser, "Nonlinear feature based classification of speech under stress," *Speech and Audio Processing, IEEE Transactions on*, vol. 9, no. 3, pp. 201–216, 2001.

[45] G. Zhou, J. Hansen, and J. Kaiser, "Methods for stress classification: nonlinear teo and linear speech based features," in *Acoustics, Speech, and Signal Processing, 1999. Proceedings., 1999 IEEE International Conference on*, vol. 4, 1999, pp. 2087–2090 vol.4.

[46] B. Priya and S. Dandapat, "Subspace filtering approach based on orthogonal projection for better analysis of stressed speech under clean and noisy environments," *International Journal of Speech Technology*, pp. 1–12, 2016. [Online]. Available: http://dx.doi.org/10.1007/s10772-016-9362-4

[47] B. Priya and S. Dandapat, "Sparse representation of lpc for analysis of stressed speech in lower dimensional subspace," in *2016 IEEE Region 10 Conference (TENCON)*, Nov 2016, pp. 661–666.

[48] B. Priya and S. Dandapat, "A subspace projection based approach to improve the recognition of stressed speech," in *2016 IEEE Annual India Conference (INDICON)*, Dec 2016, pp. 1–7.

[49] B. Priya and S. Dandapat, "Stressed speech recognition using similarity measurement on inner product space," in *Advances in Communication and Computing*. Springer, 2015, pp. 161–170.

[50] B. Priya and S. Dandapat, "Linear transformation on speech subspace for analysis of speech under stress condition," in *National Conference on Communications (NCC)*, Mumbai, India, Feb. 2015.

[51] B. Priya and S. Dandapat, "Stressed speech analysis using sparse representation over temporal information based dictionary," in *Annual IEEE India Conference (INDICON)*, Jamia Millia Islamia, New Delhi, India, Dec. 2015.

[52] D. Povey, L. Burget, M. Agarwal, P. Akyazi, F. Kai, A. Ghoshal, O. Glembek, N. Goel, M. Karafiát, A. Rastrow *et al.*, "The subspace gaussian mixture modela structured model for speech recognition," *Computer Speech & Language*, vol. 25, no. 2, pp. 404–439, 2011.

[53] S. Axelrod, V. Goel, R. Gopinath, P. Olsen, and K. Visweswariah, "Subspace constrained gaussian mixture models for speech recognition," *Speech and Audio Processing, IEEE Transactions on*, vol. 13, no. 6, pp. 1144–1160, 2005.

[54] D. Povey, L. Burget, M. Agarwal, P. Akyazi, K. Feng, A. Ghoshal, O. Glembek, N. K. Goel, M. Karafit, A. Rastrow, R. C. Rose, P. Schwarz, and S. Thomas, "Subspace gaussian mixture models for speech recognition," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, March 2010, pp. 4330–4333.

[55] G. E. Hinton, L. Deng, D. Yu, G. Dahl, A. R. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition," *Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, Nov 2012.

[56] G. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large vocabulary speech recognition," *IEEE Trans. Speech and Audio Process.*, vol. 20(1), pp. 30–42, 2012.

[57] G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schroder, "The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 5–17, Jan 2012.

[58] J. S. Park, J. H. Kim, and Y. H. Oh, "Feature vector classification based speech emotion recognition for service robots," *IEEE Transactions on Consumer Electronics*, vol. 55, no. 3, pp. 1590–1596, August 2009.

[59] E. Mower, M. J. Mataric, and S. Narayanan, "A framework for automatic human emotion classification using emotion profiles," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 5, pp. 1057–1070, July 2011.

[60] G. Fairbanks and L. W. Hoaglin, "An experimental study of the durational characteristics of the voice during the expression of emotion," *Communications Monographs*, vol. 8, no. 1, pp. 85–90, 1941.

[61] J.-C. Junqua, "The lombard reflex and its role on human listeners and automatic speech recognizers," *The Journal of the Acoustical Society of America*, vol. 93, no. 1, pp. 510–524, 1993.

[62] P. Lieberman and S. B. Michaels, "Some aspects of fundamental frequency and envelope amplitude as related to the emotional content of speech," *The Journal of the Acoustical Society of America*, vol. 34, no. 7, pp. 922–927, 1962.

[63] A. V. Oppenheim, A. S. Willsky, and S. H. Nawab, *Signals and systems*. Pearson, 2014.

[64] C. Williams, K. Stevens, and M. Hecker, "Acoustical manifestations of emotional speech," *The Journal of the Acoustical Society of America*, vol. 47, no. 1A, pp. 66–66, 1970.

[65] J.-A. Bachorowski, "Vocal expression and perception of emotion," *Current directions in psychological science*, vol. 8, no. 2, pp. 53–57, 1999.

[66] E. Lombard, "Le signe de lelevation de la voix," *Ann. Maladies Oreille, Larynx, Nez, Pharynx*, vol. 37, no. 101-119, p. 25, 1911.

[67] J. Deng, X. Xu, Z. Zhang, S. Frhholz, and B. Schuller, "Exploitation of phase-based features for whispered speech emotion recognition," *IEEE Access*, vol. 4, pp. 4299–4309, 2016.

[68] I. Luengo, E. Navas, and I. Hernaez, "Feature analysis and evaluation for automatic emotion identification in speech," *IEEE Transactions on Multimedia*, vol. 12, no. 6, pp. 490–501, Oct 2010.

[69] J. Hansen and M. Clements, "Source generator equalization and enhancement of spectral properties for robust speech recognition in noise and stress," *Speech and Audio Processing, IEEE Transactions on*, vol. 3, no. 5, pp. 407–415, Sep 1995.

[70] J. Hansen and V. Varadarajan, "Analysis and compensation of lombard speech across noise type and levels with application to in-set/out-of-set speaker recognition," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 17, no. 2, pp. 366–378, Feb 2009.

[71] H. Boril and J. H. L. Hansen, "Unsupervised equalization of lombard effect for speech recognition in noisy adverse environments," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 6, pp. 1379–1393, Aug 2010.

[72] J. H. L. Hansen, "Morphological constrained feature enhancement with adaptive cepstral compensation (mce-acc) for speech recognition in noise and lombard effect," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4, pp. 598–614, Oct 1994.

[73] J. Wagner, E. Andre, F. Lingenfelser, and J. Kim, "Exploring fusion methods for multimodal emotion recognition with missing data," *IEEE Transactions on Affective Computing*, vol. 2, no. 4, pp. 206–218, Oct 2011.

[74] M. J. Gangeh, P. Fewzee, A. Ghodsi, M. S. Kamel, and F. Karray, "Multiview supervised dictionary learning in speech emotion recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 6, pp. 1056–1068, June 2014.

[75] W. Zheng, M. Xin, X. Wang, and B. Wang, "A novel speech emotion recognition method via incomplete sparse least square regression," *Signal Processing Letters, IEEE*, vol. 21, no. 5, pp. 569–572, May 2014.

[76] M. Aharon, M. Elad, and A. Bruckstein, "K-svd: An algorithm for designing overcomplete dictionaries for sparse representation," *Signal Processing, IEEE Transactions on*, vol. 54, no. 11, pp. 4311–4322, Nov 2006.

[77] J.-J. Fuchs, "On sparse representations in arbitrary redundant bases," *Information Theory, IEEE Transactions on*, vol. 50, no. 6, pp. 1341–1344, June 2004.

[78] M. Elad, *Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing*. New York, NY, USA: Springer, 2010.

[79] Y. Chen, "Cepstral domain talker stress compensation for robust speech recognition," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 36, no. 4, pp. 433–439, Apr 1988.

[80] P. Song, Y. Jin, C. Zha, and L. Zhao, "Speech emotion recognition method based on hidden factor analysis," *Electronics Letters*, vol. 51, no. 1, pp. 112–114, 2015.

[81] H. Kathania, S. Shahnawazuddin, and R. Sinha, "Exploring hlda based transformation for reducing acoustic mismatch in context of children speech recognition," in *Signal Processing and Communications (SPCOM), International Conference on*, 2014, pp. 1–5.

[82] S. Shahnawazuddin, H. Kathania, and R. Sinha, "Enhancing the recognition of children's speech on acoustically mismatched ASR system," *Proc. IEEE TENCON*, 2015.

[83] S. Wang, Y. Zhu, L. Yue, and Q. Ji, "Emotion recognition with the help of privileged information," *IEEE Transactions on Autonomous Mental Development*, vol. 7, no. 3, pp. 189–200, Sept 2015.

[84] D. Youla, "Generalized image restoration by the method of alternating orthogonal projections," *IEEE Transactions on Circuits and Systems*, vol. 25, no. 9, pp. 694–702, Sep 1978.

[85] H. Ren and C.-I. Chang, "A generalized orthogonal subspace projection approach to unsupervised multispectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 38, no. 6, pp. 2515–2528, Nov 2000.

[86] Y. Jin, P. Song, W. Zheng, and L. Zhao, "A feature selection and feature fusion combination method for speaker-independent speech emotion recognition," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2014, pp. 4808–4812.

[87] S. Ghai and R. Sinha, "Exploring the effect of differences in the acoustic correlates of adults' and children's speech in the context of automatic speech recognition," *EURASIP J. Audio Speech Music Process.*, 2010.

[88] S. Ghai and R. Sinha, "Exploring the role of spectral smoothing in context of children's speech recognition," in *Proc. Interspeech*, 2009, pp. 1607–1610.

[89] R. Sinha and S. Ghai, "On the use of pitch normalization for improving children's speech recognition." in *Proc. Interspeech*, 2009, pp. 568–571.

[90] S. Umesh and R. Sinha, "A study of filter bank smoothing in mfcc features for recognition of children's speech," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 8, pp. 2418–2430, Nov 2007.

[91] R. Sinha and S. Umesh, "A shift-based approach to speaker normalization using non-linear frequency-scaling model," *Speech Communication*, vol. 50, no. 3, pp. 191–202, 2008.

[92] S. Shahnawazuddin, R. Sinha, and G. Pradhan, "Pitch-normalized acoustic features for robust childre's speech recognition," *IEEE Signal Processing Letters*, vol. 24, no. 8, pp. 1128–1132, Aug 2017.

[93] K. R. Scherer, "Nonlinguistic vocal indicators of 17 emotion and psychopathology," *Emotions in personality and psychopathology*, p. 495, 2013.

[94] S. Ahn and H. Ko, "Speaker adaptations in sparse training data for improved speaker verification," *Electronics Letters*, vol. 36, no. 4, pp. 371–373, Feb 2000.

[95] V. V. Digalakis and L. G. Neumeyer, "Speaker adaptation using combined transformation and bayesian methods," *IEEE Transactions on Speech and Audio Processing*, vol. 4, no. 4, pp. 294–300, Jul 1996.

[96] Y. Tang and R. Rose, "Rapid speaker adaptation using clustered maximum-likelihood linear basis with sparse training data," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 3, pp. 607–616, March 2008.

[97] C. H. Lee, C. H. Lin, and B. H. Juang, "A study on speaker adaptation of the parameters of continuous density hidden markov models," *IEEE Transactions on Signal Processing*, vol. 39, no. 4, pp. 806–814, Apr 1991.

[98] L. Welling, H. Ney, and S. Kanthak, "Speaker adaptive modeling by vocal tract normalization," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 6, pp. 415–426, Sep 2002.

[99] Y. Miao, H. Zhang, and F. Metze, "Speaker adaptive training of deep neural network acoustic models using i-vectors," *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, vol. 23, no. 11, pp. 1938–1949, 2015.

[100] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoustic, Speech and Signal Process.*, vol. 28, no. 4, pp. 357–366, August 1980.

[101] B. Yegnanarayana and R. Veldhuis, "Extraction of vocal-tract system characteristics from speech signals," *Speech and Audio Processing, IEEE Transactions on*, vol. 6, no. 4, pp. 313–327, 1998.

[102] S. Rao and W. Pearlman, "Analysis of linear prediction, coding, and spectral estimation from subbands," *Information Theory, IEEE Transactions on*, vol. 42, no. 4, pp. 1160–1178, 1996.

[103] J. Makhoul, "Linear prediction: A tutorial review," *Proceedings of the IEEE*, vol. 63, no. 4, pp. 561–580, April 1975.

[104] B. S. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *the Journal of the Acoustical Society of America*, vol. 55, no. 6, pp. 1304–1312, 1974.

[105] B. S. Atal and S. L. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave," *The Journal of the Acoustical Society of America*, vol. 50, no. 2B, pp. 637–655, 1971.

[106] H. Hermansky and N. Morgan, "Rasta processing of speech," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4, pp. 578–589, Oct 1994.

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

[108] F. Hönig, G. Stemmer, C. Hacker, and F. Brugnara, "Revising perceptual linear prediction (plp)," in *Ninth European Conference on Speech Communication and Technology*, 2005.

[109] A. Biem, S. Katagiri, E. McDermott, and B.-H. Juang, "An application of discriminative feature extraction to filter-bank-based speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 2, pp. 96–110, Feb 2001.

[110] C. Nadeu, D. Macho, and J. Hernando, "Time and frequency filtering of filter-bank energies for robust hmm speech recognition," *Speech Communication*, vol. 34, no. 1, pp. 93 – 114, 2001, noise Robust ASR. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167639300000480

[111] W.-W. Hung and H.-C. Wang, "On the use of weighted filter bank analysis for the derivation of robust mfccs," *IEEE Signal Processing Letters*, vol. 8, no. 3, pp. 70–73, March 2001.

[112] D. Reynolds, "Gaussian mixture models," in *Encyclopedia of Biometrics*. Springer, 2009, pp. 659–663.

[113] J. Niu, Y. Qian, and K. Yu, "Acoustic emotion recognition using deep neural network," in *Chinese Spoken Language Processing (ISCSLP), 9th International Symposium on*, 2014, pp. 128–132.

[114] W. L. Zheng and B. L. Lu, "Investigating critical frequency bands and channels for eeg-based emotion recognition with deep neural networks," *IEEE Transactions on Autonomous Mental Development*, vol. 7, no. 3, pp. 162–175, Sept 2015.

[115] Q. Mao, M. Dong, Z. Huang, and Y. Zhan, "Learning salient features for speech emotion recognition using convolutional neural networks," *Multimedia, IEEE Transactions on*, vol. 16, no. 8, pp. 2203–2213, 2014.

[116] R. Khosrowabadi, C. Quek, K. K. Ang, and A. Wahab, "Ernn: A biologically inspired feedforward neural network to discriminate emotion from eeg signal," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 25, no. 3, pp. 609–620, March 2014.

[117] M. Soleymani, S. Asghari-Esfeden, Y. Fu, and M. Pantic, "Analysis of eeg signals and facial expressions for continuous emotion detection," *IEEE Transactions on Affective Computing*, vol. 7, no. 1, pp. 17–28, Jan 2016.

[118] S. Shukla, S. Prasanna, and S. Dandapat, "Stressed speech processing: Human vs automatic in non-professional speakers scenario," in *National Conference on Communications (NCC)*, Jan 2011, pp. 1–5.

[119] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of german emotional speech." in *Interspeech*, vol. 5, 2005, pp. 1517–1520.

[120] J. H. Hansen, S. E. Bou-Ghazale, R. Sarikaya, and B. Pellom, "Getting started with susas: a speech under simulated and actual stress database." in *Eurospeech*, vol. 97, no. 4, 1997, pp. 1743–46.

[121] J. H. L. Hansen and B. D. Womack, "Feature analysis and neural network-based classification of speech under stress," *IEEE Transactions on Speech and Audio Processing*, vol. 4, no. 4, pp. 307–313, Jul 1996.

[122] M. Tahon and L. Devillers, "Towards a small set of robust acoustic features for emotion recognition: Challenges," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 1, pp. 16–28, Jan 2016.

[123] O. Martin, I. Kotsia, B. Macq, and I. Pitas, "The enterface' 05 audio-visual emotion database," in *22nd International Conference on Data Engineering Workshops (ICDEW'06)*, April 2006, pp. 8–8.

[124] S. Koelstra, C. Muhl, M. Soleymani, J. S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, "Deap: A database for emotion analysis ;using physiological signals," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 18–31, Jan 2012.

[125] S. S. Stevens, J. Volkmann, and E. B. Newman, "A scale for the measurement of the psychological magnitude pitch," *The Journal of the Acoustical Society of America*, vol. 8, no. 3, pp. 185–190, 1937.

[126] S. Furui, "Speaker-independent isolated word recognition using dynamic features of speech spectrum," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 34, no. 1, pp. 52–59, Feb 1986.

[127] M. Schroeder and B. Atal, "Code-excited linear prediction(celp): High-quality speech at very low bit rates," in *ICASSP '85. IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 10, Apr 1985, pp. 937–940.

[128] R. Riesz, "Differential intensity sensitivity of the ear for pure tones," *Physical Review*, vol. 31, no. 5, p. 867, 1928.

[129] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the royal statistical society. Series B (methodological)*, pp. 1–38, 1977.

[130] C. J. Wu, "On the convergence properties of the em algorithm," *The Annals of statistics*, pp. 95–103, 1983.

[131] Z. Zeng, J. Tu, B. Pianfetti, M. Liu, T. Zhang, Z. Zhang, T. S. Huang, and S. Levinson, "Audio-visual affect recognition through multi-stream fused hmm for hci," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 2, June 2005, pp. 967–972 vol. 2.

[132] L. R. Rabiner, S. E. Levinson, and M. M. Sondhi, "On the application of vector quantization and hidden markov models to speaker-independent, isolated word recognition," *Bell System Technical Journal*, vol. 62, no. 4, pp. 1075–1105, 1983. [Online]. Available: http://dx.doi.org/10.1002/j.1538-7305.1983.tb03115.x

[133] L. R. Rabiner, B.-H. Juang, S. E. Levinson, and M. M. Sondhi, "Recognition of isolated digits using hidden markov models with continuous mixture densities," *AT & T Technical Journal*, vol. 64, no. 6, pp. 1211–1234, 1985. [Online]. Available: http://dx.doi.org/10.1002/j.1538-7305.1985.tb00272.x

[134] S. Levinson, "Continuously variable duration hidden markov models for automatic speech recognition," *Computer Speech & Language*, vol. 1, no. 1, pp. 29 – 45, 1986. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0885230886800092

[135] Y.-L. Lin and G. Wei, "Speech emotion recognition based on hmm and svm," in *2005 International Conference on Machine Learning and Cybernetics*, vol. 8, Aug 2005, pp. 4898–4901 Vol. 8.

[136] B. Schuller, G. Rigoll, and M. Lang, "Hidden markov model-based speech emotion recognition," in *Multimedia and Expo, 2003. ICME '03. Proceedings. 2003 International Conference on*, vol. 1, July 2003, pp. I–401–4 vol.1.

[137] T. L. Nwe, S. W. Foo, and L. C. D. Silva, "Speech emotion recognition using hidden markov models," *Speech Communication*, vol. 41, no. 4, pp. 603 – 623, 2003. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167639303000992

[138] A. Stuhlsatz, C. Meyer, F. Eyben, T. Zielke, G. Meier, and B. Schuller, "Deep neural networks for acoustic emotion recognition: Raising the benchmarks," in *Acoustics, Speech and Signal Processing (ICASSP), IEEE International Conference on*, 2011, pp. 5688–5691.

[139] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. EPFL-CONF-192584. IEEE Signal Processing Society, 2011.

[140] J. Hershey and P. Olsen, "Approximating the kullback leibler divergence between gaussian mixture models," in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, vol. 4, April 2007, pp. IV–317–IV–320.

[141] J. Silva and S. Narayanan, "Average divergence distance as a statistical discrimination measure for hidden markov models," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 3, pp. 890–906, May 2006.

[142] G. Nakos and D. Joyner, *Linear algebra with applications*. Brooks/Cole Publishing Company, 1998.

[143] C. M. Bishop *et al.*, *Pattern recognition and machine learning*. springer New York, 2006, vol. 1.

[144] B. Varadarajan, D. Povey, and S. Chu, "Quick fmllr for speaker adaptation in speech recognition," in *Acoustics, Speech and Signal Processing. ICASSP 2008. IEEE International Conference on*, 2008, pp. 4297–4300.

[145] D. Povey and G. Saon, "Feature and model space speaker adaptation with full covariance gaussians." in *INTERSPEECH*, 2006.

[146] T. Anastasakos, J. McDonough, and J. Makhoul, "Speaker adaptive training: a maximum likelihood approach to speaker normalization," in *Acoustics, Speech, and Signal Processing. ICASSP, IEEE International Conference on*, 1997, pp. 1043–1046, vol.2.

[147] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification (2nd Edition)*. Wiley-Interscience, 2000.

[148] M. Gales, "Semi-tied covariance matrices for hidden markov models," *Speech and Audio Processing, IEEE Transactions on*, vol. 7, no. 3, pp. 272–281, May 1999.

[149] N. Kumar, "Investigation of silicon auditory models and generalization of linear discriminant analysis for improved speech recognition," Ph.D. dissertation, 1997, aAI9730738.

[150] G. Stemmer and F. Brugnara, "Integration of heteroscedastic linear discriminant analysis (hlda) into adaptive training," in *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, vol. 1, May 2006, pp. I–I.

[151] S. M. Kay, *Fundamentals of statistical signal processing*. Prentice Hall PTR, 1993.

[152] S. M. Kay, "Fundamentals of statistical signal processing: Detection theory, vol. 2," 1998.

[153] S. Young, P. Woodland, G. Evermann, and M. Gales, "The htk toolkit 3.4. 1," 2013.

[154] Y. Zhang and J. R. Glass, "Unsupervised spoken keyword spotting via segmental dtw on gaussian posteriorgrams," in *Automatic Speech Recognition & Understanding. ASRU. IEEE Workshop on*. IEEE, 2009, pp. 398–403.

[155] T. Hazen, W. Shen, and C. White, "Query-by-example spoken term detection using phonetic posteriorgram templates," in *Automatic Speech Recognition Understanding, 2009. ASRU 2009. IEEE Workshop on*, 2009, pp. 421–426.

[156] Y. Zhang and J. R. Glass, "Towards multi-speaker unsupervised speech pattern discovery," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, March 2010, pp. 4366–4369.

[157] C. F. Yeh, A. Heidel, H. Y. Lee, and L. S. Lee, "Recognition of highly imbalanced code-mixed bilingual speech with frame-level language detection based on blurred posteriorgram," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2012, pp. 4873–4876.

[158] L. F. D'Haro, R. Cordoba, M. A. Caraballo, and J. M. Pardo, "Low-resource language recognition using a fusion of phoneme posteriorgram counts, acoustic and glottal-based i-vectors," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, May 2013, pp. 6852–6856.

[159] P. R. Reddy, K. Rout, and K. S. R. Murty, "Query word retrieval from continuous speech using gmm posteriorgrams," in *2014 International Conference on Signal Processing and Communications (SPCOM)*, July 2014, pp. 1–6.

[160] A. Lee, Y. Zhang, and J. Glass, "Mispronunciation detection via dynamic time warping on deep belief network-based posteriorgrams," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, May 2013, pp. 8227–8231.

[161] H. Lee and D. Yook, "Feature adaptation for robust mobile speech recognition," *Consumer Electronics, IEEE Transactions on*, vol. 58, no. 4, pp. 1393–1398, 2012.

[162] X. Huang and K. F. Lee, "On speaker-independent, speaker-dependent, and speaker-adaptive speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 1, no. 2, pp. 150–157, Apr 1993.

[163] I. T. Jolliffe, *Principal Component Analysis*. Springer-Verlag, Berlin, Germany, 1986.

[164] S. P. Rath, D. Povey, K. Veselý, and J. Černocký, "Improved feature processing for deep neural networks," in *Proc. Interspeech*, 2013.

[165] Y. Bando, K. Itoyama, M. Konyo, S. Tadokoro, K. Nakadai, K. Yoshii, T. Kawahara, and H. G. Okuno, "Speech enhancement based on bayesian low-rank and sparse decomposition of multichannel magnitude spectrograms," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 2, pp. 215–230, Feb 2018.

[166] P. Sharma, V. Abrol, and A. K. Sao, "Deep-sparse-representation-based features for speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 11, pp. 2162–2175, Nov 2017.

[167] D. Ram, A. Asaei, and H. Bourlard, "Sparse subspace modeling for query by example spoken term detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 6, pp. 1130–1143, June 2018.

[168] S. Zhao, G. Ding, Y. Gao, X. Zhao, Y. Tang, J. Han, H. Yao, and Q. Huang, "Discrete probability distribution prediction of image emotions with shared sparse learning," *IEEE Transactions on Affective Computing*, pp. 1–1, 2018.

[169] W. Zheng, "Multichannel eeg-based emotion recognition via group sparse canonical correlation analysis," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 9, no. 3, pp. 281–290, Sept 2017.

[170] A. C. L. Ngo, J. See, and R. C. W. Phan, "Sparsity in dynamics of spontaneous subtle emotions: Analysis and application," *IEEE Transactions on Affective Computing*, vol. 8, no. 3, pp. 396–411, July 2017.

[171] R. Rubinstein, M. Zibulevsky, and M. Elad, "Efficient implementation of the k-svd algorithm using batch orthogonal matching pursuit," *CS Technion*, vol. 40, no. 8, pp. 1–15, 2008.

[172] D. L. Donoho and M. Elad, "Optimally sparse representation from overcomplete dictionaries via l1 norm minimization," *Proc. Natl. Acad. Sci. USA*, vol. 100, no. 5, pp. 2197–2002, 2003.

[173] R. Rubinstein, A. Bruckstein, and M. Elad, "Dictionaries for sparse representation modeling," *Proceedings of the IEEE*, vol. 98, no. 6, pp. 1045–1057, June 2010.

LIST OF PUBLICATIONS

Referred Journal:

1. B. Priya and S. Dandapat, “Subspace filtering approach based on orthogonal projection for better analysis of stressed speech under clean and noisy environments,” *International Journal of Speech Technology*, pp. 1−12, 2016. [Online]. Available: http://dx.doi.org/10.1007/s10772-016-9362-4

Conference Proceedings:

1. B. Priya and S. Dandapat, “A subspace projection based approach to improve the recognition of stressed speech,” in *2016 IEEE Annual India Conference (INDICON)*, Dec 2016, pp. 1−7.
2. B. Priya and S. Dandapat, “Sparse representation of lpc for analysis of stressed speech in lower dimensional subspace,” in *2016 IEEE Region 10 Conference (TENCON)*, Nov 2016, pp. 661−666.
3. B. Priya and S. Dandapat, “Stressed speech analysis using sparse representation over temporal information based dictionary,” in *Annual IEEE India Conference (INDICON)*, Jamia Millia Islamia, New Delhi, India, Dec. 2015.
4. B. Priya and S. Dandapat, “Linear transformation on speech subspace for analysis of speech under stress condition,” in *National Conference on Communications (NCC)*, Mumbai, India, Feb. 2015.
5. B. Priya and S. Dandapat, “Stressed speech recognition using similarity measurement on inner product space,” in *Advances in Communication and Computing*. Springer, 2015, pp. 161−170.