Development of Novel Speech Emotion Recognition Techniques using Machine Learning Approaches

Submitted in partial fulfilment of the requirements
for the award of the degree of

DOCTOR OF PHILOSOPHY

by

SUREKHA REDDY BANDELA

(Roll No. 716030)

Under the Supervision of
Prof. T. KISHORE KUMAR

Professor

DEPARTMENT OF ELECTRONICS AND COMMUNICATION ENGINEERING
NATIONAL INSTITUTE OF TECHNOLOGY
WARANGAL – 506004, T.S, INDIA

July – 2021

Dedicated to my beloved

Teachers, Parents and Sister

APPROVAL SHEET

This thesis entitled "**Development of Novel Speech Emotion Recognition Techniques using Machine Learning Approaches**" by **Miss. Surekha Reddy Bandela** is approved for the degree of **Doctor of Philosophy**.

Examiners

_____

_____

Supervisor

_____

Prof. T. Kishore Kumar

Professor, Electronics and Communication Engineering Department,

NIT WARANGAL

Chairman

_____

Prof. L. Anjaneyulu

Head, Electronics and Communication Engineering Department,

NIT WARANGAL

Date:

Place:

DECLARATION

I, hereby, declare that the matter embodied in this thesis entitled "**Development of Novel Speech Emotion Recognition Techniques using Machine Learning Approaches"** is based entirely on the results of the investigations and research work carried out by me under the supervision of **Prof. T. Kishore Kumar**, Department of Electronics and Communication Engineering, National Institute of Technology Warangal. I declare that this work is original and has not been submitted in part or full, for any degree or diploma to this or any other University.

I declare that this written submission represents my ideas in my own words and where other ideas or words have been included. I have adequately cited and referenced the original sources. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/date/fact/source in my submission. I understand that any violation of the above will cause for disciplinary action by the institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

Surekha Reddy Bandela

Roll No: 716030

Date:

Place: Warangal

Department of Electronics and Communication Engineering
National Institute of Technology
Warangal – 506 004, Telangana, India

CERTIFICATE

This is to certify that the dissertation work entitled **"Development of Novel Speech Emotion Recognition Techniques using Machine Learning Approaches"**, which is being submitted by Miss. Surekha Reddy Bandela (Roll No. 716030), a bonafide work submitted to National Institute of Technology Warangal in partial fulfilment of the requirement for the award of the degree of Doctor of Philosophy to the Department of Electronics and Communication Engineering of National Institute of Technology Warangal, is a record of bonafide research work carried out by her under my supervision and has not been submitted elsewhere for any degree.

Prof. T. Kishore Kumar
(Supervisor)
Professor, Department of ECE
National Institute of Technology
Warangal, India – 506004

ACKNOWLEDGEMENTS

I would like to thank a number of people who have contributed to my PhD directly or indirectly and in different ways through their help, support and encouragement.

It gives me immense pleasure to express my deep sense of gratitude and thanks to my supervisor **Prof. T. Kishore Kumar** (National Institute of Technology Warangal, NITW), for his invaluable guidance, support and suggestions. His knowledge, suggestions, and discussions helped me to become a capable researcher. He has shown me the interesting side of this wonderful multidisciplinary area and guided me to get profound knowledge as well as publications in this area.

I am thankful to the current Head of the Dept. of E.C.E., **Prof. L. Anjaneyulu**, and the former Heads, **Prof. N. Bheema Rao** and **Prof. T. Kishore Kumar** for giving me the opportunity and all the necessary support from the department to carry out my research work.

I take this privilege to thank all my Doctoral Scrutiny Committee members, **Prof. D. Srinivasacharya**, Department of Mathematics, **Dr J. Ravi Kumar**, Associate Professor, Department of Electronics and Communication Engineering and **Dr S. Anuradha**, Associate Professor, Department of Electronics and Communication Engineering for their detailed review, constructive suggestions and excellent advice during the progress of this research work.

I thank **Prof. C. B. Rama Rao** from whom I gained the knowledge on the course Advanced Digital Signal Processing as a part of my course work, which is very helpful for understanding the basics of signal processing concepts.

Special thanks to my seniors Sudeep Surendran and Sunnydayal V for their motivation, suggestions during publishing papers and being extremely supportive throughout my PhD period.

I take this opportunity to convey my regards to my speech lab-mates, NIT Warangal, Rakesh P, Prasad Nizampatnam, Ravi Bolimera, S Siva Priyanka, K Sunil Kumar for being always present next to me in time of need.

I thank my department co-scholars, M A Mushahhid Majeed and M Sandhya for being very supportive friends. I thank my hostel-mates, A Sravanthi from Dept. of Physics and D Sushmitha from Dept. of Chemical Engineering for giving me moral support throughout my PhD period.

I also appreciate the help rendered from teaching, non-teaching members and fraternity of Dept. of E.C.E. of N.I.T. Warangal. They have always been encouraging and supportive.

I acknowledge my gratitude to all my teachers and colleagues at various aspects for supporting and co-operating to complete this work.

Finally, I appreciate and respect my family members (my father Mr B. Srinivas Reddy, my mother Mrs B. Rajani and my beloved sister B. Susmitha Reddy) for being very supportive while giving me mental support and inspiration that motivated me to complete the thesis work successfully. Especially, I thank my father who has been a strong support throughout my PhD period and also helping me as an English professional to proofread my publications as well as my thesis.

SUREKHA REDDY BANDELA

ABSTRACT

Speech Emotion Recognition (SER) is defined as the process of extracting the speaker's emotional state from his or her speech. The idea of speech emotion recognition started in the early 90s mainly to detect frustration or annoyance in the speaker's voice during speech recognition system development and paved its way in many other applications.

This thesis focuses on the development of a robust speech emotion recognition system using a combination of different speech features with feature optimization techniques and speech de-noising technique to acquire improved emotion classification accuracy, decreasing the system complexity and obtain noise robustness. Novel feature fusion techniques are also developed for SER. The machine learning approach feature optimization algorithms based on feature transformation and feature selection are adopted in SER development. The feature optimization and feature selection techniques are based on unsupervised learning whereas pattern recognition or the classification of the emotions is employed using supervised learning.

The SER techniques that are initially developed used a single set of feature sets and could not achieve higher classification accuracy. The speech features are prominently divided as Continuous, Spectral, Non-linear Teager Energy Operator (TEO) and Voice Quality features. The feature fusion of these feature sets led to an improvement in speech emotion recognition accuracy rather than using a single feature set. Mel Frequency Cepstral Coefficient (MFCC), a spectral speech feature has provided promising results in SER development so far.

Even though the SER research started with the aim of detecting the stressed emotions, the emphasis is not given to stressed emotion recognition. In this thesis, a speech emotion recognition system is developed using a novel combination of TEO and spectral features for detecting stressed emotions. Spectral features - MFCC, Linear Prediction (LP) Coefficients (LPC), LP Cepstral Coefficients (LPCC) and Relative Spectral Perceptual LP (RASTA-PLP) with TEO features is used for the detection of stressed speech. The emotion recognition accuracy of the stressed emotions is improved after feature fusion.

The SER accuracy acquired using feature fusion is not up to the mark when more emotions are considered and also the combination of speech features led to a curse of dimensionality i.e., an increase in the computational overhead on the SER system. Therefore, there is a need to develop an SER system that gives better accuracy with an optimal feature set reducing the computation overhead. To select an optimized feature set, feature optimization techniques can be used to decrease the feature dimension to the most prominent one. In this thesis, a Semi-Non Negative Matrix Factorization (Semi-NMF) which is a feature transformation technique with unsupervised learning is adapted to decrease the computation overhead in the SER system.

Even though the Semi-NMF reduces the system complexity, this method lacks data interpretability. Therefore, the feature selection (FS) techniques can be used to retain the data interpretability to acquire the improved SER accuracy with reduced feature dimension. In this thesis, a novel SER system is developed using unsupervised FS techniques to reduce the huge feature set consisting of INTERSPEECH 2010 Paralinguistic features and Gammatone Cepstral Coefficients (GTCC). The FS algorithms, Unsupervised Feature Selection with Ordinal Locality (UFSOL), Feature Selection with Adaptive Structure Learning (FSASL) and a novel Subset feature selection (SuFS) technique is developed to acquire improved SER accuracy and less computational time.

The speech signal is corrupted in a noisy environment and this further causes a decrease in the SER accuracy. Therefore, a noise-robust SER system is to be developed. In this thesis, to overcome the effect of noises during speech emotion recognition, Power Normalized Cepstral Coefficients (PNCC) features that are robust to noise are used for improving SER performance. Further to obtain noise robustness in negative SNR conditions, a speech de-noising technique using NMF is adapted before SER to acquire better SER accuracy. All the developed SER systems have better performance compared to the baseline (without feature optimization/ feature selection) as well as the existing literature works. The Gaussian mixture model, k-Nearest Neighborhood, Support Vector Machine classification techniques using hold-out and cross-validation schemes are used for emotion classification in the development of the SER system.

LIST OF FIGURES

1.1 Basic Speech Emotion Recognition System……..…....6

2.1 Categorization of Speech Features……...14

2.2 Curse of Dimensionality…….……..…...19

2.3 Categorization of Feature Transformation Techniques…….....…...20

2.4 Types of Feature Selection Methods……...……..…...23

3.1 SER System with proposed Spectral and Teager feature fusion for stressed emotion recognition……..……...…...37

3.2 Nonlinear model of sound propagation along the vocal tract……..…...40

3.3 Spectral and Teager Energy Feature Fusion……..…...41

3.4 MFCC Feature Extraction……...42

3.5 LPC and LPCC Feature Extraction……..…...44

3.6 RASTA-PLP Feature Extraction……...…...47

3.7 Variation of SER accuracy for Pitch, MFCC, LPC, LPCC, MFCC-RASTA-PLP, T-MFCC, T-LPC, T-LPCC and T-MFCC-RASTA-PLP based SER system for different stressed emotions of male speakers……..…...53

3.8 Variation of SER accuracy for Pitch, MFCC, LPC, LPCC, MFCC-RASTA-PLP, T-MFCC, T-LPC, T-LPCC and T-MFCC-RASTA-PLP based SER system for different stressed emotions of female speakers...……..…….....…....54

3.9 Comparison of the proposed SER system using T-MFCC, T-LPC, T-LPCC, MFCC-RASTA-PLP, T-MFCC-RASTA-PLP, Pitch, MFCC, LPC and LPCC……..…...60

4.1 Proposed Speech Emotion Recognition System using Semi-NMF…….…..…...64

4.2 TEO-AutoCorr Feature Extraction…….....…...67

4.3 K-NN illustration……..…...72

4.4 SVM illustration…….....74

4.5 Variation of classification accuracy with Rank of Semi-NMF in the proposed SER system for MFCC, LPCC & TEO-AutoCorr Features using SVM with EMO-DB.....…...77

4.6 Variation of classification accuracy with Rank of Semi-NMF in the proposed SER system for MFCC, LPCC & TEO-AutoCorr Features using k-NN with EMO-DB.....…...78

4.7 Variation of classification accuracy with Rank of Semi-NMF in the proposed SER system for MFCC, LPCC & TEO-AutoCorr Features using SVM with IEMOCAP.....…...79

4.8 Variation of classification accuracy with Rank of Semi-NMF in the proposed SER system for MFCC, LPCC & TEO-AutoCorr Features using k-NN with IEMOCAP…. …..80

5.1 Proposed Speech Emotion Recognition system using unsupervised feature selection……...88

5.2 Variation of classification accuracy in proposed SER system with FSASL and UFSOL feature selection using SVM classifier (10-fold cross-validation) with EMO-DB…….....…102

5.3 Variation of classification accuracy in proposed SER system with FSASL and UFSOL feature selection using SVM classifier (10-fold cross-validation) with IEMOCAP……... …103

6.1 Proposed Noise Robust Speech Emotion Recognition system……115

6.2 Variation of classification accuracy in Proposed SER system with FSASL and UFSOL feature selection using SVM classifier (10-fold cross-validation) with EMO-DB……...…….....…120

6.3 Variation of classification accuracy in Proposed SER system with FSASL and UFSOL feature selection using SVM classifier (10-fold cross-validation) with IEMOCAP ……...…….……....…121

6.4 Hold-out validation accuracy variations of the proposed SER system for EMO-DB noisy data with different noises (a) airport (b) babble (c) car (d) station (e) street (f) white……...……..…126

6.5 10-Fold Cross-Validation accuracy variations of the proposed SER system for EMO-DB noisy data with different noises (a) airport (b) babble (c) car (d) station (e) street (f) white…….……....…127

6.6 Hold-out validation accuracy variations of the proposed SER system for IEMOCAP noisy data with different noises (a) airport (b) babble (c) car (d) station (e) street (f) white…….……....…128

6.7 10-Fold Cross-Validation accuracy variations of the proposed SER system for IEMOCAP noisy data with different noises (a) airport (b) babble (c) car (d) station (e) street (f) white…….....…129

6.8 Hold-out validation accuracy variations of the proposed SER system for EMO-DB after de-noising at different noisy conditions (a) airport (b) babble (c) car (d) station (e) street (f) white…….……..…132

6.9 10-fold cross-validation accuracy variations of the proposed SER system for EMO-DB after de-noising at different noisy conditions (a) airport (b) babble (c) car (d) station (e) street (f) white……..…133

6.10 Hold-out validation accuracy variations of the proposed SER system for IEMOCAP after de-noising at different noisy conditions (a) airport (b) babble (c) car (d) station (e) street (f) white……..……..…134

6.11 10-Fold Cross-Validation accuracy variations of the proposed SER system for IEMOCAP after de-noising at different noisy conditions (a) airport (b) babble (c) car (d) station (e) street (f) white…….…...…135

LIST OF TABLES

3.1 Confusion matrix of the SER system using Pitch Feature Extraction……...…...53

3.2 Confusion matrix of the SER system using MFCC and T-MFCC (proposed) feature extraction techniques……..…...53

3.3 Confusion matrix of the SER system using LPC and T-LPC (proposed) feature extraction techniques……..…...55

3.4 Confusion matrix of the SER system using LPCC and T-LPCC (proposed) feature extraction techniques……..…...55

3.5 Confusion matrix of the SER system using proposed MFCC-RASTA-PLP and T-MFCC-RASTA-PLP feature extraction techniques……...…...56

4.1 Simulation Parameters of Proposed SER System using Semi-NMF…….…...76

4.2 Comparison of Baseline and Proposed SER system with Semi-NMF for EMO-DB Database using SVM & K-NN Classifiers with Different Feature Sets…….…...81

4.3 Comparison of Baseline and Proposed SER system with Semi-NMF for IEMOCAP Database using SVM & K-NN Classifiers with Different Feature Sets…….…...82

4.4 Comparison of existing SER works with Proposed SER system for EMO-DB.……...…...83

4.5 Comparison of existing SER works with Proposed SER system for IEMOCAP…..…..…...82

5.1 INTERSPEECH 2010 paralinguistic feature set……...…...87

5.2 Simulation Parameters of Proposed SER System using unsupervised FS…….....101

5.3 Performance comparison of baseline and proposed SER systems for EMO-DB and IEMOCAP databases with SVM classifier using hold-out validation……..…….…104

5.4 Performance comparison of the baseline and proposed SER system for EMO-DB and IEMOCAP databases using SVM classifier with 10-fold cross-validation…….….105

5.5 Confusion matrix of baseline SER system for EMO-DB……...……...…106

5.6 Confusion matrix of proposed FSASL based SER system for EMO-DB……..…...…107

5.7 Confusion matrix of proposed UFSOL based SER system for EMO-DB……..…..….107

5.8 Confusion matrix of proposed FSASL-SuFS based SER system for EMO-DB...…….107

5.9 Confusion matrix of proposed UFSOL-SuFS based SER system for EMO-DB……..….108

5.10 Confusion matrix of baseline SER system for IEMOCAP……..…….……108

5.11 Confusion matrix of proposed FSASL based SER system for IEMOCAP.…… ….108

5.12 Confusion matrix of proposed UFSOL based SER system for IEMOCAP.....….108

5.13 Confusion matrix of proposed FSASL-SuFS based SER system for IEMOCAP...…...…109

5.14 Confusion matrix of proposed UFSOL -SuFS based SER system for IEMOCAP…...…109

5.15 Comparison of proposed unsupervised FS based SER system with existing works…..….110

6.1 Performance Comparison of Baseline and Proposed SER system for EMO-DB & IEMOCAP Database using SVM classifier with 10-fold Cross-Validation……..…122

6.2 Performance Comparison of Baseline and Proposed SER system for EMO-DB & IEMOCAP Database using SVM classifier with Hold-Out Validation…….….123

6.3 Performance comparison of proposed work with the existing works…….….124

6.4 Comparison of the proposed SER system using FSASL with existing works in presence of white Gaussian noise for EMO-DB database……..…130

6.5 Comparison of the proposed SER system using FSASL with existing work with noises of the Aurora database for EMO-DB and IEMOCAP databases……..…131

I Best INTERSPEECH 2010 paralinguistic features selected using UFSOL, FSASL, UFSOL-SuFS and FSASL-SuFS algorithms for the proposed SER system for EMO-DB and IEMOCAP databases……..…141

LIST OF ABBREVIATIONS

ACF – Auto Correlation Function

AMS – Amplitude Modulation Spectrogram

AR – Auto-Regressive

BBO_PSO – Biogeography Based Optimization and Particle Swarm Optimization

DCT – Discrete Cosine Transform

DFT – Discrete Fourier Transform

FFT – Fast Fourier Transform

FSASL – Feature Selection with Adaptive Structure Learning

GCZCMT – Glottal Compensation to Zero Crossings with Maximal Teager Energy Operator

GMM – Gaussian Mixture Model

GTCC – Gammatone Cepstral Coefficient

HCI – Human Computer Interaction

ICA – Independent Component Analysis

IEMOCAP – Interactive Dyadic Motion Capture

K-NN – K-Nearest Neighborhood

LDA – Linear Discriminant Analysis

LFPC – Log Frequency Power Coefficients

LLD – Low-Level Descriptor

LLE – Log-Likelihood Estimation

LPC – Linear Prediction Coefficient

LPCC – Linear Prediction Cepstral Coefficient

MDS - Multi-Dimensional Scaling

MFCC – Mel Frequency Cepstral Coefficient

NMF – Non-Negative Matrix Factorization

PCA – Principal Component Analysis

PNCC – Power Normalized Cepstral Coefficient

RASTA-PLP – Relative Spectral Perceptual Linear Prediction

RBF – Radial Basis Function

SER – Speech Emotion Recognition

SuFS – Subset Feature Selection

SNR – Signal to Noise Ratio

SSER – Stressed Speech Emotion Recognition

SVD – Singular Value Decomposition

SVM – Support Vector Machine

t-SNE – t-Distributed Stochastic Neighbor Embedding

TEO – Teager Energy Operator

TEO-CB-Auto-Env – TEO autocorrelation envelope area

TEO-AutoCorr – Teager Energy Auto-Correlation

UFSOL – Unsupervised Feature Selection with Ordinal Locality

CONTENTS

ACKNOWLEDGEMENTS i

ABSTRACT iii

LIST OF FIGURES v

LIST OF TABLES ix

LIST OF ABBREVIATIONS xii

1 Introduction 1

1.1 Introduction to Speech Emotion Recognition (SER)……..2

1.2 Applications of Speech Emotion Recognition……..……3

1.3 Basic Speech Emotion Recognition System……..……5

1.3.1 Speech Pre-Processing……..……5

1.3.2 Feature Extraction……..……6

1.3.3 Classification……7

1.4 Motivation……..……8

1.5 Problem Statement……..……8

1.6 Objectives……..……9

1.7 Organization of the Thesis……..……9

2 Literature Survey 12

2.1 Introduction…….. ……13

2.2 Speech Features……..……..13

2.3 Machine Learning for SER…… ……17

2.3.1 Feature Extraction……17

2.3.2 Feature Optimization……..……18

2.3.2.1 Feature Transformation……19

2.3.2.2 Feature Selection……..……22

2.3.3 Classification Techniques……..……25

2.3.4 Performance Metrics……26

2.4 Speech Emotion Recognition in Noisy Environments……..……28

2.5 Speech Emotion Recognition Corpora……28
2.5.1 SER corpora used in the thesis……29
2.6 Research Gaps and Issued Identified in the Development of SER system……30
2.7 Motivation for Present Work……31
2.8 Contributions……32
2.9 Summary……34

3 Speech Emotion Recognition using Spectral and Teager Energy Feature Fusion 35
3.1 Motivation……36
3.2 Proposed SER System for Stressed Emotion Recognition……36
3.2.1 Speech Pre-Processing……37
3.2.2 Feature Extraction……38
3.2.2.1 Pitch……39
3.2.3 Teager Energy Operator (TEO)……39
3.2.4 Proposed Spectral and Teager Energy Feature Fusion……41
3.2.4.1 Mel-Frequency Cepstral Coefficients (MFCC)……41
3.2.4.1(a) Discrete Fourier Transform……42
3.2.4.1(b) Mel-scale filter bank……43
3.2.4.1(c) Discrete Cosine Transform (DCT)……43
3.2.4.2 Linear Prediction Coefficients (LPC) and Linear Prediction Cepstral Coefficients (LPCC)……44
3.2.4.2(a) Auto-Correlation Function……45
3.2.4.2(b) Levinson-Durbin Algorithm……45
3.2.4.2(c) Cepstrum Analysis……45
3.2.4.3 Relative Spectral – Perceptual Linear Prediction (RASTA – PLP)……46
3.2.4.3(a) Critical Band Analysis……47
3.2.4.3(b) Equal Loudness Pre-Emphasis……48
3.2.4.3(c) Intensity Loudness Power Law……48
3.2.4.3(d) Auto-Regressive Modeling……49
3.2.5 Gaussian Mixture Model (GMM) Classifier……49
3.2.6 Hold-Out Validation……52

3.3 Simulation Results and Performance Evaluation……..……...52
3.4 Summary…….……...60

4 Speech Emotion Recognition using Semi-NMF Feature Optimization 62
4.1 Motivation…….……...63
4.2 Proposed SER System using Semi-Non Negative Matrix Factorization
(Semi-NMF)…….……...63
4.2.1 Feature Extraction…….……...64
4.2.1.1 Teager Energy Operator Auto-Correlation (TEO-AutoCorr) Feature
Extraction…….……...66
4.2.2 Feature Optimization using Semi-NMF with SVD Initialization…….……...67
4.2.3 Classification…….……...71
4.2.3.1 K-Nearest Neighborhood…….……...71
4.2.3.2 Support Vector Machine…….……...73
4.2.3.3 k-Fold Cross-Validation…….……...75
4.3 Simulation Results and Performance Evaluation…….……...75
4.4 Summary…….……...84

5 Speech Emotion Recognition using Unsupervised Feature Selection 86
5.1 Motivation…….……...87
5.2 Proposed SER system using Unsupervised Feature Selection Algorithms……..……...87
5.2.1 Feature Extraction…….……...88
5.2.1.1 INTERSPEECH 2010 Paralinguistic Feature Set…….……...89
5.2.1.2 Gammatone Cepstral Coefficients (GTCC)…….……...89
5.2.2 Unsupervised Feature Selection Algorithms…….……...90
5.2.2.1 Unsupervised Feature Selection with Ordinal Locality (UFSOL)…..……...91
5.2.2.2 Feature Selection with Adaptive Structure Learning (FSASL)……...95
5.2.2.3 Subset Feature Selection (SuFS)……...98
5.3 Simulation Results and Performance Evaluation……100
5.4 Summary…….……110

6 Noise Robust Speech Emotion Recognition using PNCC Features and NMF De-Noising **112**

6.1 Motivation……113

6.2 Proposed SER system using PNCC and NMF De-Noising……113

6.2.1 Database……114

6.2.2 Feature Extraction……114

6.2.2.1 Power Normalized Cepstral Coefficient (PNCC) Feature Extraction……114

6.2.3 Unsupervised Feature Selection……116

6.2.4 Noise Analysis……116

6.2.4.1 DenseNMF……116

6.3 Simulation Results and Performance Evaluation……119

6.3.1 Proposed SER analysis in Clean Environment……120

6.3.2 Proposed SER analysis in Noisy Environment……125

6.4 Summary……136

7 Conclusions and Future scope **138**

7.1 Conclusions……138

7.2 Future scope……140

Appendix I **141**

References **146**

List of Publications **158**

Chapter 1

Introduction

Speech emotion recognition (SER) is defined as the process of identifying the emotional state of humans from the speech signal. Emotion plays an important role in human communication in their daily lives. In the process of exchanging views between individuals, emotions reveal the state of mind of humans. During human-computer interaction (HCI), the recognition of human emotions has become vital [1]. The emotion of a person influences decision making, concentration and task solving skills. Therefore, to effectively enhance the performance of HCI, affective computing ensures that the system can recognize human emotions. This became a topic of challenge for the researchers to specialize in this domain. The different modes of emotion recognition are facial data, physiological signals, speech signal, etc. The emotion recognition popularity is widespread in many fields of application.

The scope of the thesis is to develop a speech emotion recognition system that effectively predicts the emotions by achieving higher emotion classification accuracy with less computation complexity and high noise robustness. This chapter provides a brief introduction to speech emotion recognition and the basic SER system. The motivation for

developing a novel SER system followed by problem statement, objectives, contributions and organization of the thesis is presented.

1.1 Introduction to Speech Emotion Recognition (SER):

The emotion recognition can be carried out with different data sources. Among the different modes of sources to identify emotions, the speech signal is more advantageous than biological signals such as the electrocardiogram. This is due to the fact that the speech signal can be easily acquired and economical.

Speech is the natural and fastest means of communication among humans. Thus in human-computer interaction, the idea of using the speech signal has become the most effective and fastest means of communication. Nonetheless, computers must have the ability to understand the voices of the human. From the past few decades, there is remarkable progress to make computers able to understand human speech. This process is known as speech recognition. Speech recognition is the process in which the speech signal is converted to a sequence of words. Despite the progress in speech recognition, the naturalness between human and machine is still far. This is because the machine is not able to understand the emotion of the speaker [2], [3]. To attain this, there is a need to identify the emotions from the speech signal. Due to this reason, speech emotion recognition (SER) research in this domain has been enormously increasing in the present day.

SER aims at the identification of a speaker's emotion from his or her speech. There are several emotions in human speech, depending on the various situations. Speech emotion recognition is a challenging task. The identification of the most appropriate features for differentiating emotions is difficult. This is because of the variation in the acoustics due to the differences in the speaking styles, variety of sentences spoken and speakers with different rate of speaking. These factors straight away affect the speech features like pitch and energy that are commonly used for SER [4]. The extraction of features from a speech signal to depict the emotional state of a speaker is a significant issue to be taken into consideration in a speech

recognition system. The speech features are categorized as Continuous, Spectral, Non-linear Teager Energy Operator (TEO) and Voice Quality features.

Another significant problem in SER is deciding the set of emotions that are important to classify in an automatic SER system. According to the research from the linguistic researchers, the emotional set in humans typically consists of 300 emotions [5], [6]. But classifying this huge set of emotions is extremely difficult. According to the 'palette theory', any emotion is the composition of primary emotions like the colors are a mixture of few principal colors. This theory is approved by many researchers and these emotions are primarily distinguished into six basic archetypal emotions i.e., anger, happiness, surprise, disgust, fear, sadness and neutral [7], [8].

1.2 Applications of Speech Emotion Recognition

Emotion recognition is used in several aspects of day to day applications. Emotions play a major role in human-computer interaction. The emotion recognition system aims to classify the temporal emotions of humans automatically upon receiving input speech data [1], [9]. The SER system can be used in various applications.

In Medicine, a psychiatrist needs to assess the patient's psychological state from the counselling sessions. The psychological states are whether the patient has suicidal tendencies or is under depression or even has abnormal behaviour. For such purposes, an emotion recognition system can be developed with a speech signal as the input data [10]. This can be done by training system using the speech data obtained from the counselling sessions and further identify the human emotional state. For the speech therapists, who treats the disorders of voice, speech and language, the SER system can be used as a diagnostic device. One such software is icSpeech, which records and analyzes the speech signals [11]. This is to know whether the patient is suffering from any stress. The SER system in this analysis uses prosodic, vocal tract and glottal parameters as speech feature to identify the emotional states.

The call centre services facilitate the customers to give feedback and have inquiries regarding certain products. While providing these services, the product companies have to meticulous towards providing utmost satisfaction to the customers to improve their sales. But often faces difficulties in solving the customers' conflicts. Therefore, the customer service agents have to be trained to solve the complaints with much patience. For such purposes, a system has been designed to assess customer satisfaction considering the recorded calls. However, this process takes place after attending the call. In real-time assessment, to analyze the behavioural state of the customers such as frustration, an SER system can be developed using pitch, energy and rate of speech features to detect these emotions. This improves the call attendant quality of service. Further, the system developed can also be used to detect the customers' mood to identify the urgency of situation such as panic or anger emotions and prioritize their calls. This quality will be very advantageous in emergency services to avoid mishaps.

In crime investigation, during the deception detection from the suspects, lie detectors can be very helpful. Lie Detector helps to decide if the person is speaking honest or lying. In the central bureau of investigation, the lie detector is used to find the criminals and also for avoiding corruption in the cricket council [12]. X13-VSA PRO COBRA Voice Lie Detector is a computerized software system that is advanced, sophisticated and innovative. It is a stress analyzer that identifies the truth promptly from the human voice [9].

In the banking sector, if an ATM is developed with the unique capability of combining speech recognition, speaker recognition and emotion recognition, it ensures a high level of security while infiltrating into confidential information. During the customer enrollment, the system can take their voices to ensure authenticity and also, to assess the levels of the anger, nervousness or deception signs from their speech [13]. So if there is any fraud, the ATM does not dispense the cash and rather blocks the ATM card while providing security.

In transportation, emotion recognition is mainly used to detect the frustration or stress emotions of a vehicle driver to avoid accidents [14], [15]. In carboard systems, the

information regarding the driver's emotion to provide safety arrangements such as initiating help or the system communicating with the driver to change the emotion of the driver to solve the errors [16]. In aircraft cockpits, the speech recognition systems are well equipped to identify the stress of the pilot and give commands accordingly [17].

In entertainment services such as Music Player, updates the playlist based on the mood or emotional state of the listener [18]. Speech emotion recognition can be more useful, to enhance the naturalness of the speech during the communication between the human and the computer or machine. In E-Learning and story-telling applications, the emotional state of the student or listener can be assessed in real-time and the presentation style can be adjusted accordingly [19]. In humanoid robots and robotic pets, if the robots can converse with the human understanding the emotions and also with emotions in their speech, then the conversation will be realistic and also pleasurable [20]–[22].

1.3 Basic Speech Emotion Recognition System:

The basic speech emotion recognition system consists of the pre-processing system, feature extraction and classifier blocks as shown in Figure 1.1. The physical quantities in the speech signal after the pre-processing stage are given to the feature extraction block. Here, F1, F2… Fn are the features extracted and these are given to the classifier section. Finally, a particular emotion is detected using this classifier.

1.3.1 Speech Pre-Processing:

The speech signal is pre-processed before giving it to the feature extraction module to improve the efficiency and accuracy of the feature extraction process. The pre-processing stages are Filtering, Framing and Windowing. The physical quantities like pitch, energy and formants are obtained from the speech signal after the pre-processing stage [23].

Filtering is the process used to reduce the noise in a speech signal that occurs due to disturbances in the environment or during the recording of the speech sample. The purpose of a pre-emphasis filter is to boost the energy of the speech signal in the higher frequencies which are attenuated during the speech signal production from the vocal tract.

The speech signal is not stationary and it is difficult to analyze non-stationary signals. Hence, framing the speech signal into an equal number of samples helps to analyze the signal independently. Frame size is chosen based on the feature extraction method used. An overlap between the frames is allowed to avoid the difference between the frames. When the signal is divided into frames, there exist some discontinuities at the edges of each frame of the input data signal. To avoid this discontinuity, each frame is passed through a tapered window. The various windows are Hamming, Hanning, Rectangular, Barlett, Kaiser, etc.

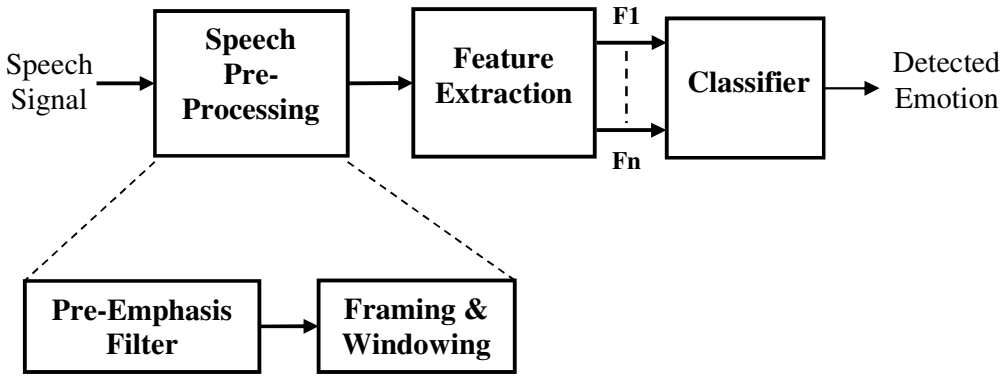


Figure 1.1: Basic Speech Emotion Recognition System

1.3.2 Feature Extraction:

Feature extraction aims to obtain the emotion relevant information from the speech signal with a reduced dimension 'n' number of features (F_1 - F_n). Identifying the speech features for emotion recognition is very important. The specific speech feature extraction techniques in speech emotion recognition help to classify the emotions from speech

efficiently. So far, many speech features have been investigated for speech emotion but the best speech feature set is not yet discovered. Speech features that are affected due to emotions are categorized as Qualitative, Spectral, Continuous and Teager Energy Operator (TEO) based features [8].

Most of the emotional content of a speech utterance affects the continuous prosody features like pitch, zero-crossing rate and energy. The speech features used under this category are related to energy, articulation rate, spectral information and fundamental frequency (f0). The perceived emotion and voice quality have a strong relation. These are classified as voice level, voice pitch, temporal and feature boundary structures. Features based on spectral analysis are shown as a short time representation of speech signal. The distribution of spectral energy of a speech utterance depends on its emotional content. It is observed that high-arousal emotions like happiness (or) anger have high energies at higher frequencies, while utterances with low-arousal emotions like sadness have less energy in a similar range. Mel Frequency Cepstral Coefficients (MFCC) and Linear Predictive Coefficients (LPC) are the most widely used spectral based features for emotion detection. Speech is produced by the non-linear airflow in the vocal tract system. The flow of air in the vocal tract system that produces sound is affected by the muscle tension of the speaker under stressful conditions. Hence, non-linear speech features are essential in the detection of speech in sound. Teager and Kaiser introduced the Teager Energy Operator feature, by considering that hearing is the manner of detection of energy. In the present day, mostly the combination of speech features is preferred in speech emotion recognition and has become a common practice.

1.3.3 Classification:

The speech features are given to the classification block to obtain the emotion of the speech signal. The classification of emotions is based on pattern recognition in Machine Learning. The classification can be either a supervised or unsupervised phenomenon. For unsupervised classification, techniques like clustering algorithms are used such as Gaussian Mixture Models (GMM), k-means clustering, etc. GMM can also be used as a supervised

classifier. For supervised classification, k-Nearest Neighborhood (K-NN), Support Vector Machine (SVM) can be used. The supervised classification techniques mostly detect the emotion accurately because of the known label or emotion information during the training of the classifier model.

1.4 Motivation

Human-computer interaction will be effective if the computer can recognize human emotions accurately. At present emotion recognition is used in various applications such as medicine, transportation, customer service, education, etc. The speech signal is one of the most useful and easily accessible data sources for emotion recognition when other sources are not available. This led to enormous research in the domain of speech emotion recognition. The idea of speech emotion recognition started in the early 90s mainly to detect frustration or annoyance in the speaker's voice and paved its way in many other applications. If stressed emotions are identified beforehand accurately, many disasters can be avoided [24]. To date, most of the research in SER has not focused on recognition of the stressed emotions. The SER system that provides accurate emotion recognition is very useful in many applications. Along with high accuracy, the system must also be real-time. But the current research on SER has not focused much on this aspect so far. Most of the existing SER systems aim only on increasing accuracy and also do not consider the noisy conditions that may arise in real-time scenarios. Due to this the existing SER systems have high computation complexity and are vulnerable to noisy conditions.

1.5 Problem Statement:

The Speech Emotion Recognition (SER) system has to be capable of identifying the emotion from the speech signal accurately. Some of the SER techniques use a single speech feature set for emotion recognition and the accuracy is <60%. The accuracy can be improved by using the combination of multiple speech features. But this increases the computational

overhead of the SER system [25], [26]. Many of the SER methods have been developed with a clean speech database till now. But, SER accuracy gets affected due to noisy environments. Hence, a speech emotion recognition system is to be developed such that it achieves higher emotion recognition accuracy with less computation complexity and is also robust to noise.

1.6 Objectives

1) Implementation of a novel Speech Emotion Recognition (SER) system using Spectral and Teager energy feature fusion for detecting stressed emotions.

2) To develop a Speech Emotion Recognition System using Semi-NMF Feature Optimization for feature dimension reduction to increase the SER classification accuracy and overcome the curse of dimensionality.

3) To develop an SER system using unsupervised feature selection algorithms to decrease the SER computational time and acquire better classification accuracy by preserving the data interpretability.

4) To develop a Noise Robust SER system by using Power Normalized Cepstral Coefficients (PNCC) and speech De-noising.

1.7 Organization of the Thesis

In **chapter 1,** the concept of speech emotion recognition and its applications are introduced. The motivation towards SER, objectives and contributions towards the thesis are discussed in brief.

In **chapter 2,** the existing literature works on SER with emphasis on the different speech features, the idea of feature fusion and feature dimension reduction using feature optimization techniques are discussed. The different SER corpora available are discussed. Also, the SER systems that are affected due to noisy environments are discussed.

In **chapter 3,** an SER system using spectral and Teager energy feature fusion for identifying stressed emotions is proposed. The spectral features are Mel Frequency Cepstral Coefficients (MFCC), Linear Prediction Coefficients (LPC), Linear Prediction Cepstral Coefficients (LPCC) and Relative Spectral Perceptual Linear Prediction (RASTA-PLP). A Gaussian Mixture Model (GMM) classifier is used for emotion classification using a hold-out validation scheme. The results of the proposed system are compared with the baseline SER system without using TEO in terms of classification accuracy.

In **chapter 4,** an SER system using semi-NMF feature optimization is proposed reducing the number of features. The features are Pitch, MFCC, Teager-Autocorrelation (TEO-AutoCorr) and MFCC+TEO-AutoCorr. The k-nearest neighbourhood (k-NN) and support vector machine (SVM) classifiers are used for emotion classification with a five-fold cross-validation scheme. The results of the SER system without feature optimization and the developed SER system with Semi-NMF are compared in terms of classification accuracy and number of features. The proposed system is also compared with existing SER works.

In **chapter 5,** a speech emotion recognition system using different feature selection techniques is proposed. The INTERSPEECH 2010 Paralinguistic features and Gammatone Cepstral Coefficient (GTCC) speech features are used in the proposed system. The Unsupervised Feature Selection with Ordinal Locality (UFSOL), Feature Selection with Adaptive Structure Learning (FSASL) and a novel Subset Feature Selection (SuFS) with the combination of UFSOL, FSASL is used for feature selection. These are unsupervised algorithms. The SVM classifier with ten-fold cross-validation and hold-out validation schemes is used for the classification of emotions. The results of the proposed system are

compared with the baseline and existing works in terms of classification accuracy and computational time.

In **chapter 6,** a noise-robust SER system is proposed using the denseNMF speech de-noising technique. The INTERSPEECH 2010 Paralinguistic features, GTCC and PNCC features are used. Before feature extraction, the denseNMF technique is used for noise removal. The SVM classifier with ten-fold cross-validation and hold-out validation schemes is used for the classification of emotions.

In **chapter 7,** the conclusions of the contributions of the thesis and the future scope of this work are discussed in brief.

Chapter 2

Literature Survey

In this chapter, the existing literature works on speech emotion recognition are discussed. Initially, the categorization of speech features, with the SER techniques developed using the single and combination of these features is presented. The drawbacks of feature fusion and the importance of feature optimization to overcome the curse of dimensionality are discussed. The different SER systems developed using feature transformation and feature selection techniques are reviewed followed by the classification techniques and performance metrics that are used for SER analysis. The SER database description and the database used in the thesis are discussed in brief.

The issues identified from the literature survey and challenges in the development of the SER system are provided. From these issues, the motivations for the research work are explained in brief.

2.1 Introduction

The research on speech emotion recognition started a few decades ago. During the development of speech recognition systems in the cockpit of the aeroplanes, the system performance is affected when the speech signal is under stressful conditions. To ensure speech recognition accuracy, many efforts are taken in the development of speech recognition systems [2], [27]. The stress in the speech signal can be due to a heavy workload or exhaustion. Also, under these stressful conditions, there are high chances of causing accidents. In this kind of situations, it is very important to identify the stress or frustration from the speech to ensure to obtain better recognition performance. The stressed speech is sometimes produced due to workload [28], [29]. This led to the interest in developing the speech emotion recognition system and research in the SER domain enormously. In implementing an SER system, speech feature extraction plays a dominant role and there is a huge variation in the emotion recognition accuracy due to the features chosen.

2.2 Speech Features

Speech features play a prominent role in the development of a speech emotion recognition system. The emotion salient information can be precisely apprehended in the speech features, these features are further used by classification or pattern recognition model for emotion detection. Speech features are majorly classified as Continuous, Spectral, Non-linear Teager Energy Operator (TEO) and Voice Quality features [7], [8]. Figure 2.1 shows the categorization of the speech.

The continuous prosodic features are pitch, zero-crossing rate, energy, formants, etc., which affect the emotional variation of a speech signal. Among all these features, the pitch has a huge variation for different emotions in humans and has been extensively used in the development of an SER system to characterize emotions [30]–[33]. The voice quality features have a strong relationship with the perceived emotion [34]. These are categorized as voice

pitch, voice level, temporal & feature boundary structures, jitter and shimmer [35], glottal waveforms and their variants [36]–[40], etc.

The Spectral features are represented as the short time representation of the speech signal. The spectral energy distribution of a speech signal varies with its emotional content. Based on this, the emotions are classified as high-arousal and low-arousal emotions. High-arousal emotions have higher energies at high frequencies viz., happiness (or) anger, whereas low-arousal emotions have less energy in the same range of frequencies viz., sadness. Compared to other speech features, spectral features were able to characterize the emotional contents more accurately [41]–[45].

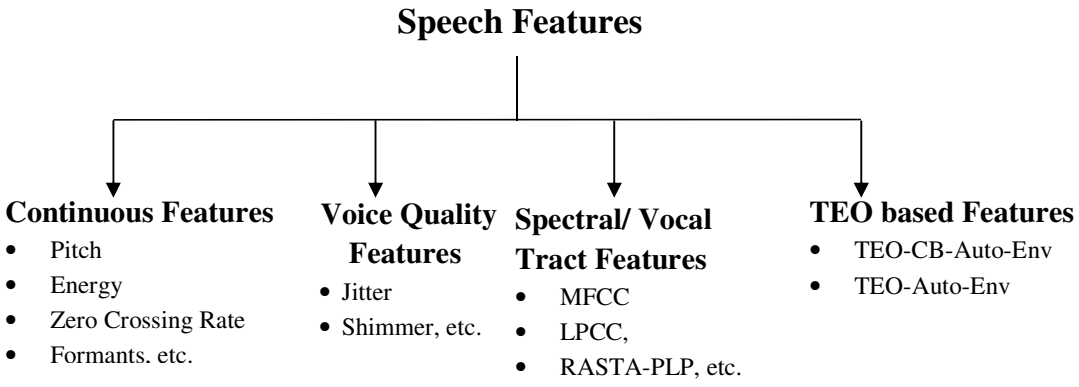


Figure 2.1: Categorization of Speech Features

It is well-known that there is a non-linear airflow during the speech production process in the vocal tract system [46], [47]. Under stressful conditions, the flow of air in the vocal tract system is affected by the muscle tension of the speaker while producing sound. These non-linear speech features are highly affected when these stressed emotional speech signals are produced. The non-linear TEO features are used for stressed emotion recognition in [48]–[50] the TEO is also combined with the glottal feature to further enhance the speech emotion recognition performance [51], [52].

Identifying the best speech feature for emotion recognition has been a difficult task for the researchers and a lot of research is carried out in this aspect. Mel Frequency Cepstral Coefficients (MFCC) [42], [43], [53]–[55], Linear Prediction Coefficients (LPC) [43], [56], Relative Spectral Perceptual Linear Prediction (RASTA-PLP) [43], variants of these features like Modified MFCC (M-MFCC) [45] and feature fusion of MFCC with Short Time Energy Features with velocity (Δ) and acceleration ($\Delta + \Delta$) [57] are some of the well-known spectral features used for speech emotion recognition. Apart from these, Log Frequency Power Coefficients (LFPC) [41], Modulation spectral features [58], Time-Frequency features with AMS-GMM Mask [59], Fourier Parameter features [60], and amplitude-based features [61] are some of the variants of spectral features that are nowadays used in speech emotion recognition analysis.

Among all these features, MFCC is the most widely used spectral feature that gave promising results in speech emotion analysis. Hence, in most of the studies, the MFCC feature set is used as a benchmark to analyze the performance of their proposed speech emotion recognition systems [41], [42], [59], [60], [62], [63]. So far, MFCC is the spectral feature providing promising results for speech emotion recognition.

For depression detection, the main focus must be on stressed emotions like anger, sadness, etc. Hence, the stressed emotion detection was started with modification and feature fusion of the different speech features. For the stressed or depressed emotion recognition, feature extraction techniques like MFCC improved as Modified MFCC (M-MFCC) [45], a new technique with feature fusion of MFCC and Short Time Energy Features with velocity (Δ) and acceleration ($\Delta + \Delta$) [53] were used. The performance of the SER systems using these feature sets is better compared to the existing MFCC and LPCC based SER systems.

Later, specifically for anger emotion recognition, acoustic (Pitch, loudness, spectral features) and linguistic (probabilistic and entropy-based words and phrases) cues [64] were introduced. Apart from these, other different feature extraction techniques like a sinusoidal model-based feature extraction technique with frequency, magnitude and phase features [65],

Empirical Mode Decomposition method with feature optimization to select particular frames of the speech signal by choosing proper filter bank [63], Hybrid Biogeography Based Optimization and Particle Swarm Optimization (BBO_PSO) by the proper selection of Higher-Order Spectral features were used for depressed emotion recognition [66].

To further improve the SER accuracy compared to MFCC based SER, the combination of qualitative and voice quality features, weighted spectral local Hu parameters are used for SER [67], [68]. In [60], [69], an SER system based on Fourier parameters, a bio-inspired Adaptive Neuro-Fuzzy Inference System technique combined with Multi-Layer Perceptron specifically to detect anger, happy and sad emotions. In [39], Glottal Compensation to Zero Crossings with Maximal Teager Energy Operator (GCZCMT) is proposed for speech emotion recognition and performed well compared to the MFCC feature. But few of the stressed emotions like anger, disgust, sadness, etc., were not accurately detected using these features also.

In [70], [71], Teager and Kaiser for the first time introduced a feature called Teager Energy Operator (TEO) to recognize the stressed emotions, by considering that hearing is the manner of detection of energy. Based on this, a Teager energy profile based pitch contour is proposed for Lombard and anger emotion recognition [48]. TEO-decomposed FM Variation (TEO-FM-Var), normalized TEO Autocorrelation envelope area (TEO-Auto-Env) and critical band based TEO autocorrelation envelope area (TEO-CB-Auto-Env) were proposed for detecting neutral versus stressed speech [50].

Low-Level Descriptors (LLDs) [72] belonging to prosodic (pitch, formants, energy, jitter, shimmer) and spectral features (spectral flux, entropy, roll-off and centroid) along with their delta and delta-delta are combined with TEO-CB-Auto-Env were designed to detect the clinically depressed/ stressed speech and it is found that TEO-CB-Auto-Env +delta +delta-delta combined with formants, log energy+ delta+ delta-delta, shimmer+ delta, spectral flux and spectral roll-off feature technique provided the highest accuracy among all the combinations for depressed emotion recognition. But this method becomes very complex due

to the presence of the combination of many features. Later various feature fusion methods with a combination of glottal, prosodic, spectral and TEO based features were proposed for stressed emotion recognition [73]–[75].

It is evident from the literature, that the combination of speech features i.e., feature fusion increases the classification accuracy of the SER system and has become the most common practice in this field [34], [57], [61].

The INTERSPEECH Emotion Challenge set with 384 features is one of the famous feature fusion sets that is used for SER [57], [76]. This feature set consists of 16 low-level descriptors and their corresponding 12 functionals. This feature set is further extended by adding some more features to incorporate the paralinguistic information assessment, the INTERSPEECH Paralinguistic Challenge set [77] with 1582 features is the recently used feature fusion set in SER for achieving better accuracy [78]–[81].

2.3 Machine Learning for SER

Machine Learning is the study of computer algorithms that improve automatically through experience [25]. At the present day, machine learning has become highly promising in solving the problems related to the field of signal processing. Among the wide range of areas in machine learning, feature extraction, feature optimization and classification are adapted in this research work for the task of speech emotion recognition.

2.3.1 Feature Extraction

The process of converting raw data such as a speech signal into adaptable data that can be processed is feature extraction. The raw data has a larger dimension and requires a lot of computational resources for processing. Feature extraction enables the reduction of the data

dimension by obtaining a set of features effectively. The obtained feature set describes by embedding the information regarding the original data precisely. This feature extraction speeds up the computational steps in the machine learning process. The different speech feature extraction techniques that are used for speech emotion recognition are discussed in section 2.2.

2.3.2 Feature Optimization

Even though the classification accuracy of the SER system increases due to feature fusion, the computational overhead also increases on the classifier. The reason is - only some of the features are useful for SER analysis whereas many other features have no role in emotion recognition. And using these irrelevant speech features decreases the performance of the SER system and leads to the curse of dimensionality as shown in figure 2.2. It is evident from figure 2.2, after a particular feature dimension threshold, the SER accuracy is decreased with the increase in the dimension of the feature set. So, by choosing an appropriate feature dimension, optimal performance can be achieved. The feature optimization methods simplify the task of choosing the optimal feature set. These techniques majorly eradicate the loss caused due to the curse of dimensionality and also solve the problem of overfitting by improving the generalization in the model, i.e., the use of less redundant data for SER that leads to incorrect predictions, thereby increasing the classification accuracy and enhancing the prediction performance by decreasing the computational time and memory used by the system.

Also, the increase of speech features will result in the increase of the computational complexity and may cause the over-fitting problem, i.e., the model achieves better accuracy while training, but it fails while testing on new data [82], [83]. These drawbacks can be overcome by adapting feature dimension reduction techniques before the classification of the feature sets.

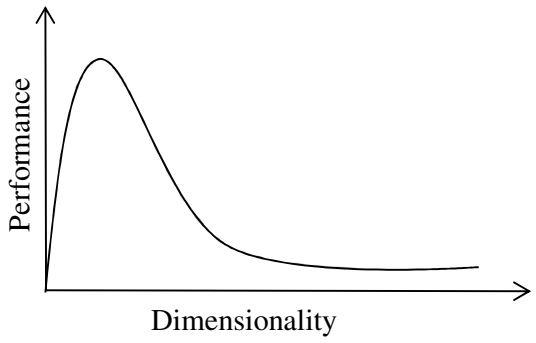


Figure 2.2: Curse of Dimensionality

Therefore, it is preferred to reduce the number of features by performing feature selection or optimizing the feature set before emotion classification. There are several feature selection and optimization techniques for dimension reduction of the feature set to overcome the disadvantages of having huge feature sets. In the feature selection, a subset of the original features is selected, which retain the desired feature set. In machine learning, a feature vector is an *n*-dimensional vector representing the features of all samples. The space related to these vectors is the feature space. To decrease the dimensionality of feature space, feature selection or feature transformation methods can be used. In feature transformation, the original feature space is transformed into a different space having a distinct set of axes to reduce the dimensionality of the data. The discriminant feature information is concentrated in a particular part of the coefficients in the transformed domain. In feature selection, the significant features are chosen rather than transforming to another domain. Several feature selection techniques were used by the researchers to select the more appropriate feature set [84], [85].

2.3.2.1 Feature Transformation

Dimensionality Reduction of the feature vector is the simplest and direct way to solve the problem of high dimensionality. However, there can be uncertain data loss and reduction in classification accuracy causing instability in the SER system because of reducing the number of feature vectors. This problem can be overcome using linear transformation techniques. Consider an n-dimensional input feature vector $x = [F_1, F_2, \ldots, F_n]^T$ and then the

transformed output vector will be $y = [b_1, b_2, \ldots, b_r]^T, where\ r \ll n$ and 'r' is the reduced dimension.

For this purpose, in machine learning, several optimization techniques for dimensionality reduction are being developed to obtain the best relevant/ optimized feature set to improve the speech emotion recognition accuracy. These techniques are classified based on labelling the feature data i.e., as supervised or unsupervised. In supervised techniques the feature datasets are labelled and whereas in unsupervised techniques, the datasets are not labelled [26], [83]. These supervised and unsupervised techniques are further classified based on feature transformation into linear and non-linear techniques, in which the high dimensional feature sets are scaled down to a lower-dimensional space preserving the locality and geometric structures.

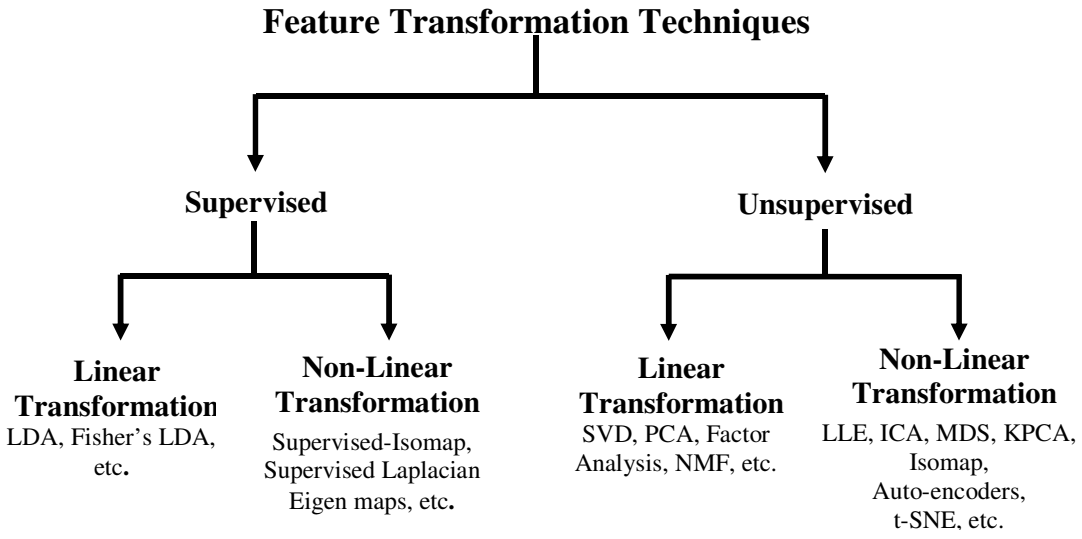


Figure 2.3: Categorization of Feature Transformation Techniques

In Linear Transformation based dimension reduction, the structure of a given dataset and its internal relationships are determined using Euclidean distance. Some of these are

Linear Discriminant Analysis (LDA), Fisher's LDA, etc in case of supervised and Singular Value Decomposition (SVD), Principal Component Analysis (PCA), Factor Analysis, NMF etc., under unsupervised that are based on second-order statistics and they use covariance matrix for transformations. Whereas, the Non-linear Transformation technique recover the useful and meaningful submanifolds from high dimensional datasets [83]. It also helps to understand and visualize the recovered submanifolds of complex real-time datasets Techniques. Supervised Isomap, Supervised Laplacian Eigen Maps, etc come under supervised nonlinear transformation and Log-Likelihood Estimation (LLE), Independent Component Analysis (ICA), Multi-Dimensional Scaling (MDS), Kernel PCA (KPCA), Isomap, Auto-encoders, t-Distributed Stochastic Neighbor Embedding (t-SNE), etc are some of the unsupervised nonlinear transformation techniques.

Principal Component Analysis (PCA) is one of the important & most used dimensionality reduction technique that is based on feature transformation for speech emotion recognition where the feature data is transformed from high dimensional feature space to a lower dimension [33], [84]. Many other feature optimization techniques viz., Singular Value Decomposition (SVD) [82], Locally Linear Embedding (LLE) [86], Non-Negative Matrix Factorization (NMF) [87] and Linear Discriminant Analysis (LDA), a supervised machine learning technique [33], are few other feature optimization techniques used for speech emotion recognition. In SVD and NMF, the complete set of features transforms with matrix factorization, to obtain a lower-dimensional feature set, acquiring an optimal feature set. In [88], [89], the variants of autoencoders namely adversarial and variational autoencoders are used to transform the huge feature sets into lower dimension and this reduced feature set is used for speech emotion recognition to acquire high performance.

In [90], semi-NMF feature transformation technique with multiple kernel Gaussian process and in [91], a supervised feature transformation method i.e., modified supervised locally linear embedding (MSLLE) are used as feature dimension reduction techniques in SER.

2.3.2.2 Feature Selection

In feature selection, from the original feature set, a subset of features is selected with respect to their relevance and redundancy. It improves the prediction performance and reduces computational complexity and storage, providing faster and cost-effective models [92].

In feature selection, the original feature space is reduced into a subspace without transformation. Some examples of feature selection methods are ReliefF, Fisher Score, Information Gain, Chi Squares, LASSO, etc. Feature selection techniques can be categorized based on the labelling of the data as supervised, unsupervised and semi-supervised. In supervised feature selection, the data is labelled as a feature evaluation process, whereas, if the data is huge, the labelling of the data is costly and also a tedious task. Unsupervised feature selection can overcome these drawbacks of supervised approaches. But this is more difficult than supervised ones since it does not have labelled data and still, its result can be good even without any prior knowledge. The evaluation of feature selection methods can be further classified into four types, i.e., filter, wrapper, embedded, hybrid and ensemble feature selection, as depicted in figure 2.4.

Filter feature selection techniques use statistical analysis to assign a distinctive feature with a score. Their score ranks the features, and later, these are retained or removed from the original feature vector set accordingly. These filter techniques mostly use a single variable in their analysis and features are considered independent of each other or sometimes dependent terms. The most commonly used filter methods are the Chi-squared test[93], variance threshold [94], information gain, etc. The fast feature selection method, i.e., Fisher feature selection is used in [95] with decision SVM for SER.

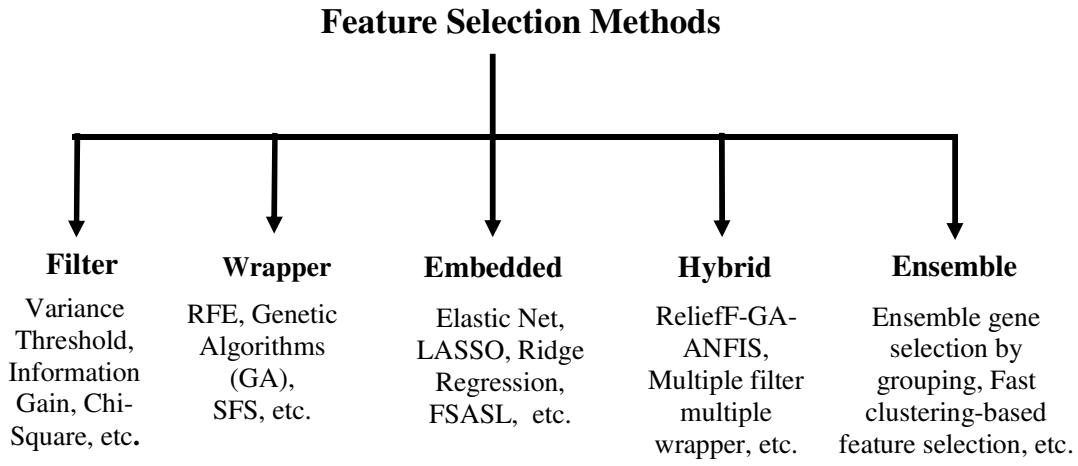


Figure 2.4: Types of Feature Selection Methods

The wrapper feature selection techniques consider a set of features with various combinations of the feature subsets. Later, these feature subsets are compared with one another as a search problem which is estimated and compared with other groups. Further, the prediction process is performed to assign the score onto each of the feature sets depending on the prediction accuracy. The search process can be systematic, stochastic or heuristics such as the best-first search, random hill-climbing algorithm, forward and backward passes to add and remove features. Genetic algorithms, Recursive Feature Elimination (RFE), Sequential Feature Selection (SFS), etc. are some of the wrapper methods of feature selection. In [96], SFS and Sequential Floating Feature Selection (SFFS) are used for SER.

Embedded methods, in the learning process, select the features that are best to improve accuracy. The most frequently used feature selection embedded methods are regularization techniques. In [97], for SER multiple kernel learning based on L1-Norm and embedded feature selection is used.

The Hybrid method is a combination of two or more feature selection methods (e.g., filter + wrapper). These methods try to acquire the benefits of both techniques by combining

their corresponding strengths. It achieves improved efficiency, prediction performance and decreases computational complexity. The most widely used hybrid method is the combined feature selection with filter and wrapper approaches.

The ensemble method constructs a collection of feature subgroups and produces an aggregate result from the group. The primary goal of this method is to tackle the unpredictability problems in most feature selection algorithms. This method is based on various subsampling schemes in which one feature selection technique runs on many subsamples, and the resultant features are combined to attain a subset with more stability. With this, for high dimensional data, the feature selection performance is no longer dependent on any individual selected subset, thus attains more flexibility and robustness.

In [98], the SER system uses feature selection based on sparse representation, i.e., sparse partial least squares regression (SPLSR). Apart from these feature selection techniques, feature transformation methods can also be used in SER for the reduction in feature dimension [90], [91], [99]. In [100], unsupervised feature learning is carried out using k-means clustering, sparse autoencoders (AE) and sparse restricted Boltzmann machines for feature mapping to obtain optimal feature set for SER. The adversarial AEs and variational AEs can encode the high dimensional feature vector to a lower dimension and also have the ability to reconstruct the original feature space. Therefore, in [88], [101], these are used as feature dimension reduction techniques for SER.

In [89], a new variant of feature extraction technique i.e., deep neural network based heterogeneous model consisting of AE, de-noising AE and an improved shared hidden layer AE is used to extract the features from the speech signal. These layers also provide feature optimization up to some extent. But to obtain better performance for SER with the high-dimension feature set, a fusion level network with a support vector machine (SVM) classifier is used.

2.3.3 Classification Techniques

Classification is the process of recognizing the pattern of a certain set of categorical data. The classification algorithms use features as the input data for classifying the emotions in SER [102]. The learning of the data in the classifier can be either supervised or unsupervised. In supervised learning, initially, the training data is labelled with the corresponding emotion label during the training of the classifier and further, the unlabeled test data is predicted. Whereas in unsupervised learning the training data is not labelled and the training, as well as testing of the data, is performed.

Some of the classification techniques used in the detection of emotions after the feature extraction process are the Hidden Markov Model (HMM), GMM, Neural Network, k-NN, SVM, etc. After the feature extraction, different classification models like Support Vector Machine (SVM), Vector Quantization (VQ), Gaussian Mixture Model (GMM), k-Nearest Neighborhood (k-NN) and different Neural Network (NN) algorithms are used for identifying particular emotion from the features extracted [7], [8], [25].

In HMM, the classifier is physically linked to the speech production process, has been in extensive use in applications of speech like speech emotion recognition, isolated word recognition, speech recognition and speech segmentation. The HMM is a stochastic process that contains a first-order Markov chain whose states are hidden. The random process related to each state generates an observation sequence. These hidden states are captured using the temporal structure of the data.

GMM is a probabilistic model to estimate the density of multivariate normal densities. This is a special case of continuous HMM containing only a single state. As the testing and training requirements are very few compared to HMM, these are effectively used to model the multimodal distributions. GMMs are very useful for the extraction of global features in speech emotion recognition from training data. GMMs assume that all their vectors of training

and testing are independent of each other. Therefore, the temporal features are not effectively modelled using GMMs. The most favourable number of Gaussian components required to model the GMM is a complex task. Model order section criteria are the most common method used to find the number of Gaussian components.

Artificial Neural Network (ANN) is another classifier used for speech emotion recognition. They are mostly used for modelling nonlinear mappings. ANNs are categorized into three main basic types: Multilayer Perceptron (MLP), Radial Basis Function (RBF) and Recurrent Neural Networks (RNN) networks. MLP has a well-defined training algorithm and can be easily implemented once the ANN structure is specified. There are many design parameters to be considered in ANNs like defining many hidden layers, number of neurons and neuron activation function in each layer.

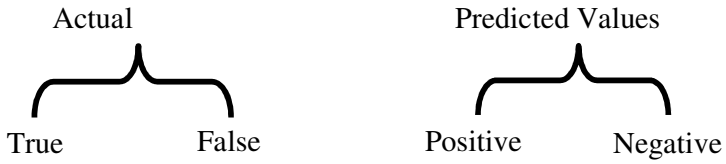
2.3.4 Performance Metrics

The performance metric used in machine learning for classification is the confusion matrix. The metrics that can be obtained from the confusion matrix are accuracy, recall, specificity, precision and ROC curve. If a classification task is considered to classify two or more categories, then the confusion matrix consists of four combinations of actual and predicted values as shown in table 2.1.

Table 2.1: Confusion Matrix

		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	True Positive (TP)	False Positive (FP)
	Negative (0)	False Negative (FN)	True Negative (TN)

The actual value is the category for the target variable in the dataset, whereas the predicted value is what the machine learning model predicts the test data to be. The value of the target variable can be specified to be Positive and Negative. The corresponding binary value of positive and negative can be 1 and 0 respectively. For example, if a classification problem is chosen to predict whether a patient has a disease or not. Positive with binary '1' signifies that patient 'has a disease' whereas the negative indicates the 'patient has no disease'. True positive is prediction is correct to be positive, true negative implies the prediction is that the result is correct to be negative, false positive and false negative are the errors, indicating that the prediction is wrong to be positive or negative. In short, the positive and negative are exemplified in accordance with the predicted values, while the True and False using actual values.



The classification accuracy that is used in SER to validate the system performance can be measured from the confusion matrix. In the emotion classification task, the categories are emotion classes. Accuracy metric gives the best understanding of the classification or prediction performance of a system intuitively i.e., the percentage of the predicted classes that are identified correctly. This is measured as the ratio of true values with the total number of occurrences i.e.,

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+FN+TN} \tag{2.1}$$

2.4 Speech Emotion Recognition in Noisy Environments

In a real-time scenario, the SER system becomes vulnerable to some of the unwanted noisy conditions. There is a need to make the SER system robust to noisy conditions. In [14], an adaptive speech enhancement technique with three-level wavelet packet decomposition is adopted for de-noising the noisy speech in the presence of highway, parking lot and city street noises recorded on Infinity Q45 test-bed along with white Gaussian noise, before SER. In most SER research work, Mel-Frequency Cepstral Coefficients (MFCC) is proven to be the most useful and widely used feature extraction technique [42], [59], [103], [104]. In [104], Teager energy based MFCC feature extraction is used to achieve noise robustness towards white noise in SER. In [105], an SER system, robust to white noise is proposed using an enhanced sparse representation classifier i.e., a weighted sparse representation model based on maximum likelihood estimation. In [106], Power Normalized Cepstral Coefficients (PNCC) feature extraction method consisting of asymmetric noise suppression and weight smoothing to acquire noise robustness in SER in the presence of babble, white, speech shaped and factory (Noisex-92) noises. In [107], multiple-kernel learning with sample reconstruction of noise is used for noise robustness of SER in the presence of white noise. In [108], a sub-band spectral centroid wavelet packet cepstral coefficients with importance weighted SVM classifier is used for robust SER against white noise. In [109], an SER system with the combination of MFCC, pitch and MFCC derived from wavelet-based speech features is proposed for robustness to different noises of Aurora noisy dataset [110] and white Gaussian noise.

2.5 Speech Emotion Recognition Corpora

Speech emotion recognition is evolved into one of the challenging tasks in the speech processing domain. The performance of the SER system in the real-time environment depends on the naturalness of the speech signal. Hence, choosing the appropriate speech database for SER system development is a significant issue. The database with lower quality affects the

classification of emotions and the system might give incorrect emotion predictions. While developing a database on speech emotion, the factors like the language, gender of the speaker, number of subjects, emotion type, age, etc. are to be considered.

Speech corpora used for developing SER systems can be divided into 3 types i.e., Actor (simulated) based, Elicited (induced) and Natural (Spontaneous) emotional speech databases. The acted database is collected from experienced and trained theatre or radio artists incorporating the aspects of relevant emotions. It is fully developed in nature and typically intense. This is also known as a full-blown emotional database. Elicited emotion database is recorded by simulating a situation without knowledge of the speakers. The speakers are involved in an emotional conversation and this database is more natural. The natural emotion database is mildly expressed. It is difficult to recognize the emotions in it. This type of database is generally recorded from call centre conversations, cockpit recordings, a dialogue between patients or between doctor and patient. Some of the well-known emotion databases that are widely used in SER research are EMO-DB, IEMOCAP, eNTERFACE, EMOVO, SAVEE, BAUM-1s challenge database, EMA, etc.

2.5.1 SER corpora used in the thesis:

The speech corpora used in most of the SER works are EMO-DB and IEMOCAP databases. Hence, these databases are used for evaluating the performance of the proposed algorithms in this thesis.

EMO-DB is the most prominently used German database in SER research work [111]. The recording for emotional data is done in an anechoic chamber by five male and five female actors between the age group of 25-35. In a recording environment at 48 kHz, 535 speech signals are recorded comprising of anger, anxiety/ fear, happiness, boredom, disgust, sad and neutral emotions. Further, in this SER analysis, these speech signals are down-sampled to 16 kHz.

The Interactive Emotional Dyadic Motion Capture (IEMOCAP) database is an acted, multimodal and multi-speaker database [112]. Twelve hours of audio-visual data that include video, speech, text transcriptions and motion capture of the face. In this work, the speech data with emotions, anger, happiness, neutral and sadness are considered as in most of the SER works, with a total of 4490 utterances.

2.6 Research Gaps and Issues Identified in the Development of SER System

From the literature survey on speech emotion recognition, few research gaps need to be addressed for further improving the emotion classification accuracy:

1) The existing speech emotion recognition systems use a single set of features rather than using multiple feature combinations and have less SER accuracy <60% [8], [34], [61].

2) The stressed emotions like anger, anxiety, could not be detected efficiently using the existing and most widely used MFCC or LPC based SER systems [50].

3) In most of the speech emotion recognition systems, the accuracy is increased by combining number features. While, this increases the computational complexity of the overall system that is the curse of dimensionality [25], [113]. The feature optimization techniques can serve the purpose of decreasing the feature dimension without causing any loss in the data.

4) The feature optimization in SER can be accomplished either by feature transformation or feature selection [57], [114]. But in speech emotion recognition, using the transformation techniques for feature optimization, the entire feature set is transformed into a new dimension. Due to this, there is a lack of data interpretability and the transformation becomes expensive with huge data [95], [115].

5) The pre-processing of speech data into frames gives better SER accuracy. Even though, increasing the frame size in speech pre-processing for SER, there is no decrease in emotion recognition accuracy [116]. But still, the existing SER systems use a lower frame size. But by using a lower frame size in SER, the feature data dimension is high and increases computational complexity.

6) In real-time scenarios, the speech signal becomes vulnerable in presence of noisy conditions and this causes reduction in SER accuracy. Speech de-noising techniques such as sparse representation are used for developing a noise-robust SER [105]. But still, the emotion recognition accuracy is very less compared to the SER system in a clean speech scenario.

2.7 Motivation for Present Work

1) It is evident from the existing SER works, by using the combination of multiple speech features the speech emotion recognition accuracy can be increased enormously. Hence, better combinations of speech features need to be adapted in the development of the SER system for acquiring higher accuracies.

2) TEO based speech features have been specifically developed for stress emotion recognition. Therefore, an SER system with TEO speech features with a better combination of other speech features can be developed for stressed speech emotion recognition.

3) To address the issue of the curse of dimensionality, feature optimization techniques can be used for the reduction of feature data dimension. The feature optimization is performed in two types i.e., feature transformation and feature selection. Any of these methods have to be considered based on the data, such that there is no information loss and also less computation time for system processing.

4) For optimizing the feature data without the loss of data interpretability, feature selection methods can be used. In feature selection, the original feature space is reduced into a subspace without transformation. The features that provide the utmost information about emotions can be chosen using the feature selection algorithms. The unsupervised algorithms further reduce the work of labelling the dataset and hence, can be most ideal for feature optimization.

5) Since there is the same accuracy comparably for both the SER systems with lower and higher frame size. The computation time in SER can be further decreased by increasing the frame size in the speech pre-processing stage with which the data dimension is reduced.

6) The SER system in real-time environments must be robust to noise to provide better emotion recognition accuracy. This can be achieved by speech de-noising of the noisy speech signal before performing the emotion recognition.

2.8 Contributions

In the first contribution, the speech emotion recognition system is developed using the feature fusion of spectral and Teager energy feature fusion for detecting stressed emotions. TEO is designed for increasing the energy levels of stressed emotions like anger, anxiety, etc. The TEO feature combined with spectral features gives better stressed emotion detection rather than using the individual spectral features. The spectral features are Mel-Frequency Cepstral Coefficients (MFCC), Linear Prediction Coefficients (LPC), Linear Prediction Cepstral Coefficients (LPCC) and Relative Spectral – Perceptual Linear Prediction (RASTA-PLP). The Gaussian mixture model (GMM) through hold-out validation is used for emotion classification.

In the second contribution, an SER system using the Semi-NMF dimension reduction technique is proposed. In SER, the combination of speech features improves the performance of the SER system but this result in an increase in the computational complexity. The feature

optimization using Semi-Non Negative Matrix Factorization (Semi-NMF) with Singular Value Decomposition (SVD) initialization technique is proposed to reduce the number of speech features acquiring comparably better recognition accuracy. The SER system is initially developed using MFCC, LPCC, TEO-Auto correlation (TEO-AutoCorr) and a combination of these feature sets. These speech features are optimized using the Semi-NMF algorithm. The k-Nearest Neighborhood (K-NN) and Support Vector Machine (SVM) with 5-fold cross-validation are used for emotion classification.

In the third contribution, an SER system is developed using unsupervised feature selection algorithms. The accuracy of the SER system is increased by adopting Semi-NMF optimization. This is a feature transformation technique, where the feature data is transformed into a new domain and thus there is no data interpretability. Also, this transformed data is not reliable as it tends to decrease the emotion recognition accuracy in noise environments. Therefore, an SER system is proposed using unsupervised feature selection algorithms UFSOL, FSASL and SuFS to select the best features to reduce the computational complexity and also retaining the data interpretability acquiring comparable accuracy. Further, rather than using low frame size in speech pre-processing, higher frame size is used to reduce the data dimension and thus decreasing the computation time in the SER system. The speech features used are INTERSPEECH 2010 paralinguistic and Gammatone Cepstral Coefficient (GTCC). The SVM classifier is used to classify the emotions using the selected speech features from feature selection.

In contribution 4, a noise-robust SER system is developed using PNCC features combined with INTERSPEECH and GTCC features with denseNMF speech denoising technique in the pre-processing stage. The SER using unsupervised feature selection algorithms is robust to noises with SNR levels higher than 15dB for the EMO-DB database and in the case of the IEMOCAP database for SNR levels higher than 10dB. Specifically, in the presence of babble noise, the performance of the SER system decreases further. Therefore, the DenseNMF denoising is adopted in the SER system before feature extraction to acquire noise robustness.

2.9 Summary

In this chapter, the origin of speech emotion recognition and the review of the different types of speech features used for SER are discussed. The advantage and drawbacks of using speech feature fusion for improving classification accuracy are discussed. The curse of dimensionality issue and the feature optimization phenomenon for solving this problem is explained. The machine learning concepts: feature extraction, feature optimization with an emphasis on feature transformation and feature selection followed by classification for SER are reviewed. The SER database description and the databases used in the thesis are presented in brief. The issues identified from the literature survey, the motivation towards this research work and the contributions of the thesis are emphasized

Chapter 3

Speech Emotion Recognition using Spectral and Teager Energy Feature Fusion

In this chapter, the combination of spectral features with Teager Energy Operator (TEO) is proposed for the detection of stressed emotions from the speech signal. TEO is specifically designed for increasing the energies of the stressed speech signals whose energy is reduced during the speech production process. The spectral features considered are Mel Frequency Cepstral Coefficients (MFCC), Linear Prediction Coefficients (LPC), Linear Prediction Cepstral Coefficients (LPCC) and Relative Spectral –Perceptual Linear Prediction (RASTA-PLP). Speech emotion recognition system using the feature extraction techniques namely, a spectral feature fusion method MFCC-RASTA-PLP and Spectral + Teager Energy-based features using the combination of the spectral features with TEO are proposed. In this analysis, the Emo-DB database is used with four stressed emotions namely anger, fear, disgust and sad along with neutral. The Gaussian Mixture Model (GMM), an unsupervised classifier is used for emotion classification.

3.1 Motivation

Majorly, the detection of stressed emotions in speech plays a vital role in the real-time applications like the mood of a car driver to avoid accidents, a student's mental state to give them proper counselling, a child's psychological state to improve their parents and other acquaintance, etc. and also in aircraft cockpits the speech recognition systems trained with stressed speech provided better results compared to normal speech [24]. If these stressed emotions are identified beforehand then it is possible to avoid many disasters happening. To date major research was focused on recognition of all the emotions and emphasis was not given to the stressed emotions like anger, fear, sadness, disgust, frustration, etc. Therefore, the TEO feature that is specifically designed for detecting the stressed emotions can be used for stressed speech emotion recognition. As the feature fusion increases the emotion recognition, the combination of TEO with spectral features is used in the proposed system.

3.2 Proposed SER System for Stressed Emotion Recognition

Figures 3.1 shows the proposed speech emotion recognition system used for the detection of four different stressed emotions namely, anger, fear, disgust and sad with reference to the neutral speech. Finally, a GMM classifier with supervised learning is used to classify these emotions. In this proposed SER system for the stressed emotion analysis, the combination of spectral features - MFCC, LPC, LPCC and RASTA-PLP features with TEO are used. Compared to the rest of the speech features, spectral features have been widely used for SER and hence, used in the proposed system along with TEO. The baseline SER system is developed using the pitch and spectral features individually i.e., without combination with the TEO feature set. And further, these results are compared with the proposed SER system to ensure the importance of the TEO feature.

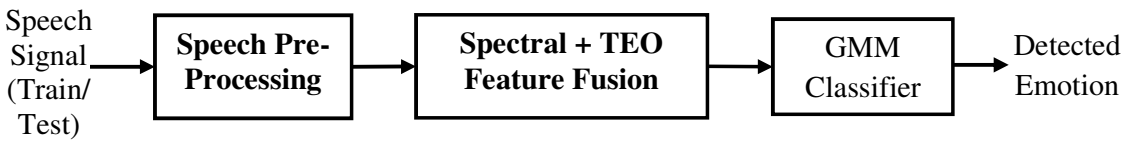


Figure 3.1: SER System with proposed Spectral and Teager feature fusion for stressed emotion recognition

The speech signal is passed through the Teager Energy block to enhance the energies of the stressed or stressed emotional contents of the speech and further, the energized signal is given to the Spectral Feature Extractor.

3.2.1 Speech Pre-Processing

Initially, the speech signal is pre-processed to improve the efficiency of the feature extraction process and the stages in pre-processing are Filtering, Framing and Windowing [23]. The pre-emphasis filter performs as a first-order high pass filtering that is used to boost the energy of the speech signal in the higher frequencies which are attenuated during the speech signal production from the vocal tract. If $s[n]$ is considered as a speech signal, then the time domain and z-domain representations of a pre-emphasis filter are given as [23]:

$$h[n] = s[n] - \alpha s[n-1] \quad \text{(or)} \quad H(z) = S(z)[1 - \alpha z^{-1}] \tag{3.1}$$

Here 'α' is the filter coefficient and its value must be between 0.9 and 1. The value of 'α' is chosen to be 0.97 in the proposed system. It is well-known that the speech signal is not a stationary signal and hence, it is difficult to analyze the speech signals. To overcome this problem, the pre-emphasized speech signal is framed into an equal number of samples. Now, each frame can be individually considered stationary and the signal processing techniques can be applied.

Each frame consists of an equal number of samples and it is termed as Frame Length. The number of frames varies from one speech signal to another signals depending on the length of the speech signal. When the signal is divided into frames, there exist some discontinuities at the edges of each frame of the input speech signal. To avoid this discontinuity, each frame is passed through a tapered window. There are different types of windows used in speech pre-processing like Hamming, Hanning, Barlett, etc. An overlap between the frames must be allowed so that there is no loss in the speech signal information. Hamming window is chosen in this work, as it provides less spectral leakage at the edges of the frames. The window size is chosen based on the frame length ('N'). The Hamming window is given by [23]:

$$w[n] = 0.54 - 0.46\cos(2\pi\frac{n}{N}), \text{ where } 0 \leq n \leq N \tag{3.2}$$

Here, 'N' is the window size, 'n' is the speech signal length. In this work, the frame length is considered as '256' and the overlap allowed between the frames is chosen to be '80'. Later, the pre-emphasized speech signal '$h[n]$' is multiplied with this window function by allowing a frame overlap to obtain the resultant signal as,

$$x[n] = h[n] * w[n] \tag{3.3}$$

3.2.2 Feature Extraction

In SER, the process of extracting emotion specific information from the raw speech signal is known as Feature extraction. In this proposed SER system 24 MFCC, 21 LPC, 21 LPCC and 21 RASTA-PLP coefficients are extracted from one speech signal. The pitch and the proposed combination of spectral with TEO feature fusion are discussed below.

3.2.2.1 Pitch:

Pitch is the speech signal's fundamental frequency that is the reciprocal of the fundamental period that gives pitch value. The high or low frequency of a sound is due to the variation in the pitch of a speech signal. The pitch can be estimated directly from the waveform using an Autocorrelation function. The autocorrelation function is:

$$R(l) = \frac{1}{N}\sum_{i=0}^{N-l-1}(x[i+l]x[i]);\ l \geq 0 \tag{3.4}$$

Here, the input speech signal is $x[i]$, time of the discrete signal is 'i' and 'l' is the delay introduced. $R(l)$ has a higher value, if $x[i]$ is equal to $x[i+l]$. The rest of the feature extraction techniques considered are discussed in the subsequent section.

3.2.3 Teager Energy Operator (TEO)

The speech under stressful conditions affects the nonlinear flow of air in the vocal tract system when the speech signal is produced. Hence, these non-linear speech features are very important for the detection of speech.

Teager proposed an energy operator i.e., a measure of speech signal energy based on his experiments known as the Teager Energy operator [48]. In the experiments, Teager showed that the flow of air in the vocal tract is separated and then follows the vocal tract walls. Based on the observations made on the results of a few whistle experiments, Teager proposed the vocal tract geometry and modelled the speech production as shown in Figure 3.2. In this model, air exits the glottis as a jet and attaches to the nearest wall of the vocal tract. When the air is passed between the true and the false vocal folds through the cavity, the vortices of air are formed. During the propagation of air, most of it is passed near the lips following the vocal tract walls. The important portion of this model is the action of the vortex. In the traditional speech production model, the sound is produced actively in an unconstructed vocal tract only

at the glottis. Whereas, Teager stated that vortices in the region of the false vocal folds also produce sound actively and this causes modulations in the speech signal.

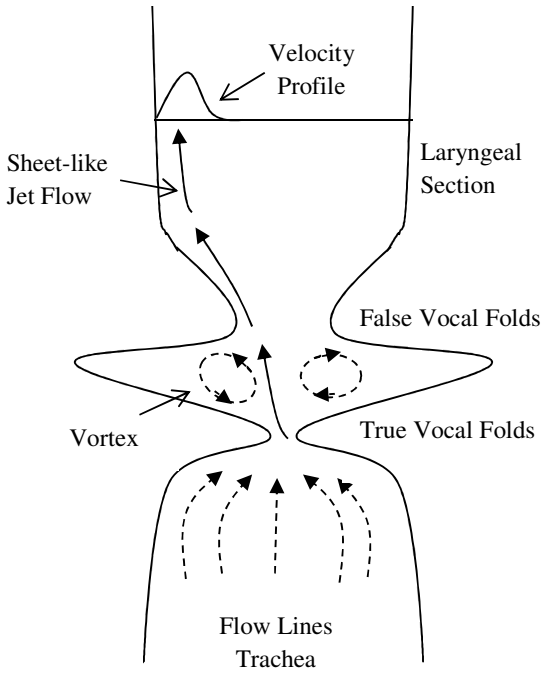


Figure 3.2: Nonlinear model of sound propagation along the vocal tract [48]

Later, Teager conducted several experiments on the hearing process and came up with a measurement of energy parameter to find proof of speech modulation patterns. J. Kaiser in [71], for the first time, showed the energy operator as follows,

$$TEO\big(x(t)\big) = \left(\frac{d}{dt}x(t)\right)^2 - x(t)\left(\frac{d^2}{dt^2}x(t)\right) \text{ (or) } TEO(x[n]) = x^2[n] - x[n]x[n+1] \tag{3.5}$$

where $x(t)$ and $x[n]$ are the continuous and discrete speech signals.

TEO is used for detecting stressed emotions like Lombard, angry, loud versus neutral emotions.

3.2.4 Proposed Spectral and Teager Energy Feature Fusion

The combination of TEO before spectral feature extraction is the proposed spectral and Teager energy feature fusion technique. The spectral features that are combined with TEO are MFCC, LPC, LPCC and RASTA-PLP, this is shown in figure 3.3.

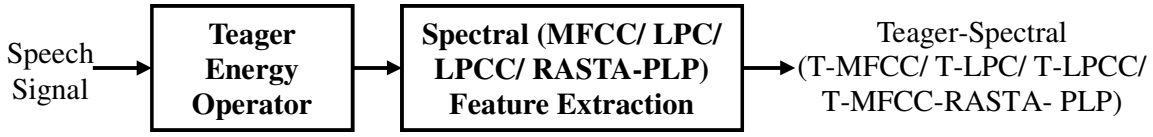


Figure 3.3: Spectral and Teager Energy Feature Fusion

3.2.4.1 Mel-Frequency Cepstral Coefficients (MFCC) Feature Extraction

According to human perception, the frequency of the speech signal is according to the Mel-scale. MFCC features are extracted from the disintegration of triangular filter-bank spaced on according to the mel-scale. MFCC is one of the popular spectral transformation techniques used in speech recognition and also in speech emotion recognition. It mimics the perception of a human ear, using cepstral analysis [117]. MFCCs are computed by dividing the speech into frames and the illustration of MFCC is as shown in figure 3.4. Initially, the speech signal is pre-processed that includes pre-emphasis, framing and windowing as discussed in section 3.2.1. Then these speech frames are fed to discrete Fourier transform to transform the speech signal into the frequency domain.

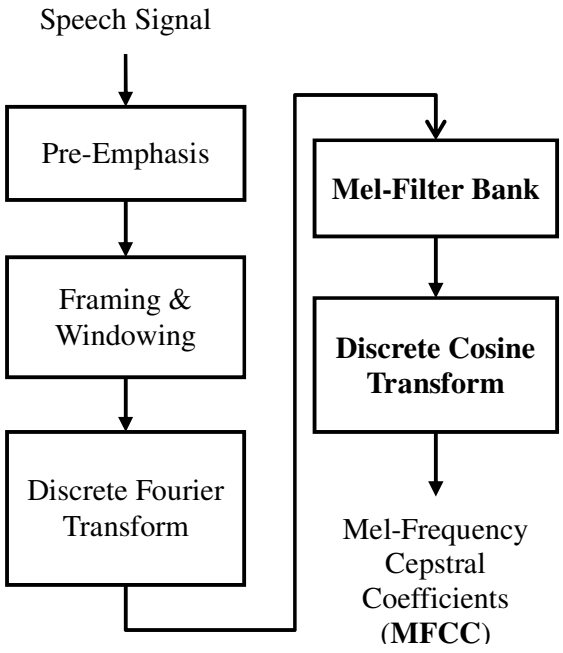


Figure 3.4: MFCC Feature Extraction

3.2.4.1(a) Discrete Fourier Transform:

After pre-processing spectral information of the speech signal has to be extracted. The Discrete Fourier transform (DFT) transforms the windowed speech segment into the frequency domain. DFT is used for the extraction of spectral information of a discrete-time signal in discrete frequency bands. A commonly used algorithm for computing the DFT is the Fast Fourier Transform (FFT).

$$X(k) = \sum_{i=0}^{M-1} \left(x(i) e^{-j2\pi k \frac{i}{M}} \right); \quad (3.6)$$

Here, 'x(i)' is input sequence with $i = 0,1,2, \ldots \ldots, M-1$, 'M' is the number of points in DFT and $k = 0,1,2, \ldots \ldots, M-1$.

With an increase in the number of DFT points the frequency resolution increases but the information is diluted by the same factor. Hence, the solution is the use of a more optimum number of points in the frequency spectrum, such that much of the speech signal information is utilized.

3.2.4.1 (b) Mel-scale filter bank:

The FFT spectrum obtained has a very wide frequency range and the speech signal does not vary on a linear scale. To compute '*Mel*' for a given frequency '*f*' in Hertz,

$$mel(f) = 2595 \log_{10}(1 + \frac{f}{700}) \quad \text{(or)} \quad mel(f) = 1127 \ln(1 + \frac{f}{700}) \tag{3.7}$$

Mel-frequency is directly proportional to the log of linear frequency. It has a linear spacing of frequencies below 1 kHz and logarithmic above 1 kHz. A series of triangular band-pass filters are used in a mel-scale to find the weighted sum of the spectral components of the filter such that the output signal estimates as a Mel-scale.

3.2.4.1 (c) Discrete Cosine Transform (DCT):

Mel Frequency Cepstral Coefficients are obtained by converting the Mel spectrum to the time domain as shown in [74]. And hence, each input utterance is converted into the sequence of acoustic vectors. Here, DCT is applied to the log energy 'E_k' and the triangular band pass filters are used to obtain '*L*' MFCCs. DCT is given by,

$$C_{\mathrm{m}} = \sum_{k=1}^{N} \left(\mathrm{E_k} \cos(\mathrm{m}(\mathrm{k} - \tfrac{1}{2}) \tfrac{\pi}{\mathrm{N}}\right); \tag{3.8}$$

where, *m*=1, 2… *L*; '*N*' is the number of log spectral coefficients and '*L*' is the Mel-scale Cepstral Coefficients obtained.

3.2.4.2 Linear Prediction Coefficients (LPC) and Linear Prediction Cepstral Coefficients (LPCC) Feature Extraction

The speech signal is the convolution of the excitation source and time-varying vocal tract system components. To analyze these components independently the signal has to be separated and hence, Linear Prediction (LP) analysis is used to find the source and system components from the time domain. In this analysis, the linear combination of the past time-domain samples, $s[n-1], s[n-2], \ldots, s[n-M]$ to predict the current sample $\hat{s}[n]$ in time-domain:

$$\hat{s}[n] = -\sum_{k=0}^{M} \mathrm{a}_k \mathrm{s}[\mathrm{n} - \mathrm{k}] \tag{3.9}$$

where $\mathrm{a}_k, \mathrm{k} = 1,2,3, \ldots\ldots, \mathrm{M}$ are called the predictor (or) LPC coefficients. Autocorrelation and Levinson Durbin algorithms are used to find the LP coefficients.

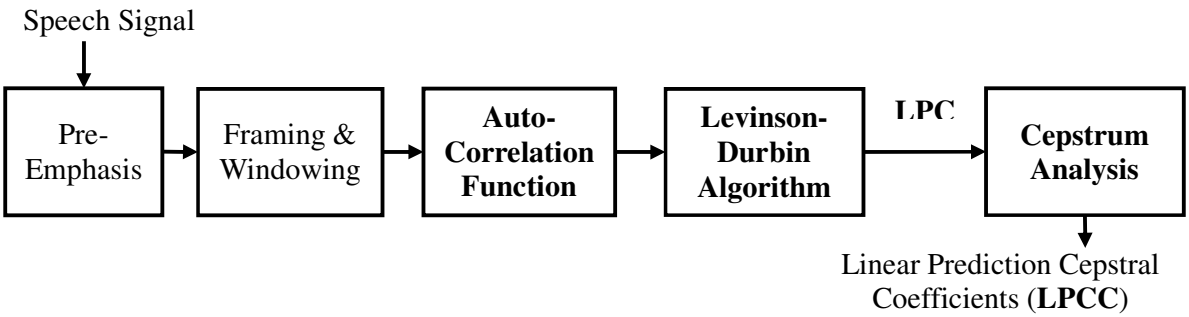


Figure 3.5: LPC and LPCC Feature Extraction

From the existing literature studies, it is proved that the combination of various features improves the accuracy of the system [53]. Hence, Cepstrum analysis is combined with LP analysis to acquire the LPCC. The feature extraction of LPC and LPCC is shown in figure 3.5.

3.2.4.2(a) Auto-Correlation Function (ACF):

ACF is the correlation of a signal with a delayed copy of itself as a function of delay. Informally, it is the similarity between observations as a function of the time lag between them [118].

$$R_{xx}(k) = \sum_{n=k}^{M-1} \mathrm{s[n]s[n-k]} \tag{3.10}$$

where 's[n]' is the input signal to the function and $'k'$ is the delay parameter.

3.4.2.2(b) Levinson-Durbin Algorithm:

The Levinson-Durbin recursion is an algorithm for finding an all-pole IIR filter with a prescribed deterministic autocorrelation sequence. It has applications in filter design, coding, and spectral estimation. The filter coefficients produced using the Levinson algorithm have a minimum phase [118].

$$H(z) = \frac{1}{A(z)} = \frac{1}{1+a(2)z^{-1}+\cdots+a(n+1)z^{-n}} \tag{3.11}$$

3.2.4.2(c) Cepstrum Analysis:

The cepstral coefficients are derived from the LPC coefficients derived using the following recursion [119]:

$$C_0 = \log_e p$$

$$C_m = a_m + \sum_{i=1}^{m-1} \frac{i}{m} C_i a_{m-i} \text{ , for } 1 < m < p$$

$$C_m = \sum_{i=m-p}^{m-1} \frac{i}{m} C_i a_{m-i} \text{ , for } m > p \quad (3.12)$$

The resultant $\{ C_0, C_1, \ldots\ldots, C_m \}$ are LPCC features. The LPCC feature extraction is designed to obtain '21' features and is used in this analysis.

3.2.4.3 Relative Spectral – Perceptual Linear Prediction (RASTA-PLP) Feature Extraction:

RASTA-PLP feature extraction technique uses RASTA filtering in Perceptual Linear Prediction (PLP). PLP coefficients [75] are created from the LP coefficients by performing perceptual processing i.e. critical band analysis, equal loudness pre-emphasis and intensity loudness before performing the Auto-Regressive (AR) modelling. The feature extraction is shown in figure 3.6. RASTA Filtering was introduced along with PLP, i.e. using the bandpass filter in the log spectral domain [120], [121]. By using this, the slow variations in the channel are suppressed.

A general RASTA filter is defined by:

$$T(z) = \frac{k \sum_{n=o}^{N} (n - \frac{n-2}{2}) z^n}{1 - p x^{-1}} \quad (3.13)$$

where the numerator is a regression filter of N^{th} order and the denominator is an integrator.

The first block in any feature extraction method is pre-processing of the speech signal as discussed in section 3.2.1, the real and imaginary components of this short-term speech spectrum are squared and added to get the short-term power spectrum [75]:

$$P(\omega) = Re[S(\omega)]^2 + Im[S(\omega)]^2] \tag{3.14}$$

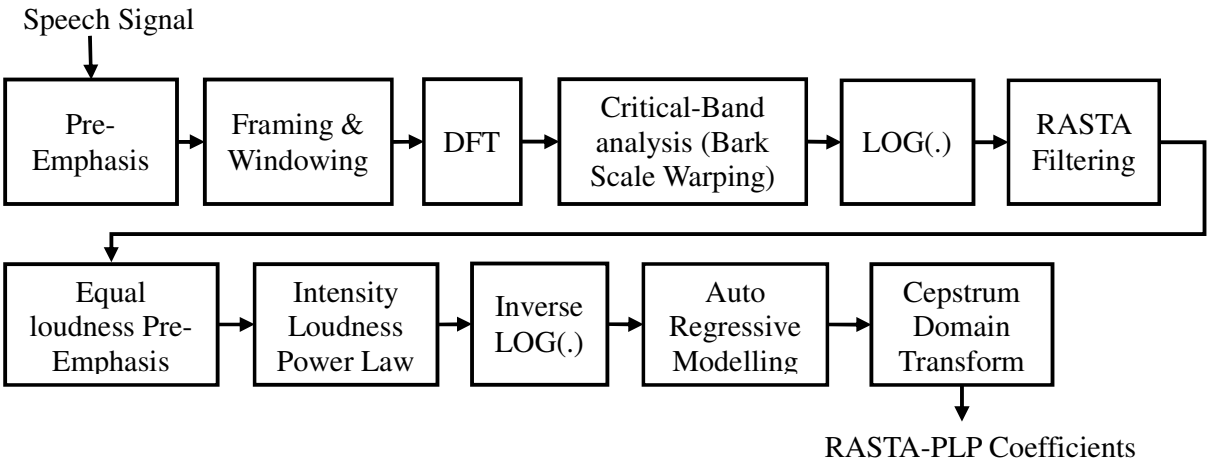


Figure 3.6: RASTA-PLP Feature Extraction

3.2.4.3(a) Critical Band Analysis:

The spectrum $P(\omega)$is warped along its frequency axis ω into the Bark frequency Ω by

$$\Omega(\omega) = 6\ln\{\frac{\omega}{1200\pi} + [\left(\frac{\omega}{1200\pi}\right)^2 + 1]\}^{0.5} \tag{3.15}$$

where ω is the angular frequency in rad/s.

This warped power spectrum is convolved with the power spectrum of the simulated critical-band curve $\psi(\Omega)$. $\psi(\Omega)$ is given by

$$\psi(\Omega) = \begin{cases} 0 & \text{for } \Omega < -1.3, \\ 10^{2.5(\Omega+0.5)} & \text{for } -1.3 < \Omega < -0.5 \\ 1 & \text{for } -0.5 < \Omega < 0.5 \\ 10^{-1.0(\Omega-0.5)} & \text{for } 0.5 < \Omega < 2.5 \\ 0 & \text{for } \Omega > 2.5 \end{cases} \quad (3.16)$$

The discrete convolution of $\psi(\Omega)$ with $P(\omega)$yields samples of the critical-band power spectrum as:

$$\Theta(\Omega_i) = \sum_{\Omega=-1.3}^{2.5} P(\Omega - \Omega_i)\psi(\Omega) \quad (3.17)$$

The convolution with the relatively broad critical-band masking curves $\psi(\Omega)$ significantly reduces the spectral resolution of $\Theta(\Omega)$ in comparison with the original $P(\omega)$. This allows for the down-sampling of $\Theta(\Omega)$.

3.2.4.3(b) Equal Loudness Pre-Emphasis

Later, the sampled $\Theta[\Omega(\omega)]$ is filtered using RASTA filtering by following equation (3.18) and the resulted signal $T(\Theta[\Omega(\omega)])$ is pre-emphasized by the simulated equal-loudness curve

$$\Xi[\Omega(\omega)] = E(\omega)T(\Theta[\Omega(\omega)]) \quad (3.18)$$

$E(\omega)$ is an approximation to the non-equal sensitivity of human hearing at different frequencies and simulates the sensitivity of hearing at about the 40 dB level.

3.2.4.3(c) Intensity Loudness Power Law:

Before finding the all-pole model, cubic root amplitude compression is performed as:

$$\phi(\Omega) = \Xi(\Omega)^{0.33} \tag{3.19}$$

This gives an approximation to the power law of hearing and simulates the nonlinear relation between the intensity of sound and its perceived loudness. This operation reduces the spectral-amplitude variation of the critical-band spectrum so that the following all-pole modelling can be done with a lower model order.

3.2.4.3(d) Auto-Regressive Modeling:

Finally, $\phi(\Omega)$ is approximated by the spectrum of an all-pole model using the autocorrelation method and further, the autoregressive coefficients could be transformed into cepstral coefficients of the all-pole model.

3.2.5 Gaussian Mixture Model (GMM) Classifier

GMM is a probabilistic model to estimate the density of multivariate normal densities. GMMs are very useful for the extraction of global features in speech emotion recognition from training data. GMMs assume that all their vectors of training and testing are independent of each other. Therefore, the temporal features are not effectively modelled using GMMs. The most favourable number of Gaussian components required to model the GMM is a complex task. Model order section criteria are the most common method used to find the number of Gaussian components.

The features extracted for each emotion of male and female speakers of the EMO-DB database need to be clustered to differentiate different emotions. This clustering can be done using a GMM classifier. A collection of independent Gaussian distributions are produced by GMMs. In this, each data point corresponds to each of these distributions or clusters.

A training set $\{x(1), \ldots\ldots, x(m)\}$ is considered in this model. The data is modeled using a joint distribution $p(x(i), z(i)) = p(x(i)|z(i))p(z(i))$. Here z(i) is the multinomial ($\emptyset$) with $\emptyset_j \geq 0$. where, $\emptyset_j \geq 0$, $\sum_{j=1}^{k} \emptyset_j$=1 and the parameter $\emptyset_j$ gives $p(z(i) = j)$ and $x(i)|z(i) = j \sim N(\mu_j \sum_j)$. The number of values that the z(i) can take on is denoted by k. Thus, the hypothesis is that by randomly choosing $z(i)$ from $\{1, \ldots, k\}$ each $x(i)$ was generated. Depending on $z(i)$ $x(i)$ is drawn from one of the k Gaussians. This is the mixture of the Gaussian model. The z(i)'s are hidden or unobserved random variables. Therefore, the parameters of the model are thus $\emptyset, \mu \; and \sum$. To estimate these values, the likelihood of the data is:

$$N(\emptyset, \mu, \sum) = \sum_{i=1}^{m} logp(x(i); \emptyset, \mu, \sum) = \sum_{i=1}^{m} log \sum_{z(i)=1}^{k} p(x(i)|z(i); \mu, \sum) p(z(i), \emptyset) \quad (3.20)$$

Specifically, the log-likelihood is:

$$N(\emptyset, \mu, \sum) = \sum_{i=1}^{m} logp(x(i)|z(i); \mu, \sum) + logp(z(i); \emptyset) \quad (3.21)$$

When this likelihood is maximized with respect to $\emptyset, \mu \; and \sum$, then the following parameters are obtained:

$$\emptyset_j = \frac{1}{m} \sum_{i=1}^{m} 1\{z(i) = j\} \quad (3.22)$$

$$\mu_j = \frac{\sum_{i=1}^{m} 1\{z(i)=j\}x(i)}{\sum_{i=1}^{m} 1\{z(i)=j\}} \tag{3.23}$$

$$\sum_j = \frac{\sum_{i=1}^{m} 1\{z(i)=j\}(x(i)-\mu_j)(x(i)-\mu_j)^T}{\sum_{i=1}^{m} 1\{z(i)=j\}} \tag{3.24}$$

If the $z(i)$'s are known, then maximum likelihood estimation is nearly similar to the estimated parameters of the Gaussian discriminant analysis model. But, the $z(i)$'s are not known in the density estimation problem. The EM algorithm is an iterative algorithm with two steps. In the E-step, the values of the $z(i)$'s are guessed. In the M-step, based on these guesses, the parameters of the model are updated. The maximization becomes easy in the M-step it is pretended that the guesses in the first part are correct. The algorithm is:

Repeat until convergence:

(i) E-step: For each i, j, set

$$w_j(i) \coloneqq p(z(i) = j|x(i); (\emptyset, \mu, \sum)) \tag{3.25}$$

(ii) M-step: Update the parameters:

$$\emptyset_j := \frac{1}{m}\sum_{i=1}^{m}\{w_j(i)\} \tag{3.26}$$

$$\mu_j = \frac{\sum_{i=1}^{m} w_j(i)x(i)}{\sum_{i=1}^{m} w_j(i)} \tag{3.27}$$

$$\sum_j = \frac{\sum_{i=1}^{m} w_j(i)(x(i)-\mu_j)(x(i)-\mu_j)^T}{\sum_{i=1}^{m} w_j(i)} \tag{3.28}$$

In the E-step, the posterior probability of the parameters of the $z(i)$'s is calculated. If the $x(i)$ given and if the current setting of our parameters is used, then from Bayes rule:

$$p(z(i) = j|x(i); (\emptyset, \mu, \sum)=\frac{p(x(i)|z(i)=j;\mu,\sum)p(z(i)=j;\emptyset)}{\sum_{l=1}^{k} p(x(i)|z(i)=l;\mu,\sum)p(z(i)=l;\emptyset)} \quad (3.29)$$

where $p(x(i)|z(i) = j; \mu, \sum)$ estimates the Gaussian density with mean μ_j and covariance $\sum_j$.

Here, $p(x(i)|z(i) = j; \mu, \sum)$ is given by evaluating the density of a Gaussian with mean μ_j and covariance Σ_j at $x(i)$; $p(z(i) = j; \emptyset)$ is given by $\emptyset_j$, and so on. In the M-step the updates are contrasted when the $z(i)$'s were known exactly.

3.2.6 Hold-Out Validation:

Hold-out is where the dataset is split up into a 'train' and 'test' set. The training set is what the model is trained on, and the test set is used to see how well that model performs on unseen data. A common split when using the hold-out method is using 80% of data for training and the remaining 20% of the data for testing.

3.3 Simulation Results and Performance Evaluation

The proposed SER system is carried out using Emo-DB Berlin German Emotional database considering the stressed emotions anger, fear, disgust and sadness along with neutral. An overlap between the frames is allowed so that there is no loss in the speech signal information. In the proposed system, the frame length is considered to be 256 and the overlap allowed between the frames is chosen as 80. The Gaussian mixture model is unsupervised during the training process with 5 clusters i.e., each cluster signifies one emotion and later,

these clusters are labelled accordingly with their corresponding emotions. During the testing phase, these emotion labels are used for predicting the test input in the proposed SER system.

Tables 3.1 to 3.5 show the confusion matrices with classification accuracies of different emotions for speech emotion recognition system for both male and female speakers using pitch, MFCC and T-MFCC, LPC and T-LPC, LPCC and T-LPCC, MFCC-RASTA-PLP and T-MFCC-RASTA-PLP i.e., existing and proposed feature extraction techniques.

Table 3.1: Confusion matrix of the SER system using Pitch Feature Extraction

Classification accuracy (%)						
Emotion	**Gender**	**Anger**	**Fear**	**Disgust**	**Neutral**	**Sad**
Anger	Male	**93.3**	0	6.7	0	0
	Female	**60**	40	0	0	0
Fear	Male	40	**20**	20	13	7
	Female	33	**40**	27	0	0
Disgust	Male	0	20	**80**	0	0
	Female	6.7	6.7	**66.6**	13	7
Neutral	Male	0	28.4	0	**46.6**	25
	Female	0	0	20	**66.6**	13.4
Sad	Male	0	0	0	40	**60**
	Female	0	0	7	0	**93**

From table 3.1, using an emotion recognition system based on pitch feature extraction, the highest accuracy is achieved for the anger emotion as 93.3%, next for disgust emotion as 80% and the lowest is for fear emotion i.e., 20%. Likewise, for the female speech, the highest accuracy is achieved for sad emotion, the next same accuracy for disgust and neutral with 66.6% and further the lowest accuracy is for fear emotion with 40% accuracy. Also, it can be inferred that mostly the fear emotion is identified as anger emotion with 40% accuracy for male and female speech.

Table 3.2: Confusion matrix of the SER system using MFCC and T-MFCC (proposed) feature extraction techniques

Classification accuracy (%)											
Emotion	**Gender**	**Anger**		**Fear**		**Disgust**		**Neutral**		**Sad**	
		MFCC	**T-MFCC (Proposed)**	**MFCC**	**T-MFCC (Proposed)**	**MFCC**	**T-MFCC (Proposed)**	**MFCC**	**T-MFCC (Proposed)**	**MFCC**	**T-MFCC (Proposed)**
Anger	Male	**100**	**100**	0	0	0	0	0	0	0	0
	Female	**93.3**	**93.3**	0	0	6.7	6.7	0	0	0	0
Fear	Male	0	0	**86.7**	**100**	0	0	13.3	0	0	0
	Female	13.3	6.7	**73.3**	**80**	6.7	6.7	6.7	0	0	6.7
Disgust	Male	0	0	26.7	0	**60**	**100**	13.3	0	0	0
	Female	6.7	0	0	0	**86.6**	**93.3**	0	6.7	6.7	0
Neutral	Male	0	6.7	0	0	0	0	**93.3**	**80**	6.7	13.3
	Female	0	0	0	0	13.3	0	**86.7**	**100**	0	0
Sad	Male	0	0	6.7	6.7	0	0	13.3	6.7	**80**	**86.6**
	Female	0	0	0	0	0	0	6.7	6.7	**93.3**	**93.3**

From table 3.2, the accuracy of the emotion recognition system for male speech using MFCC provides 100% accuracy for anger emotion and using T-MFCC 100% accuracy is achieved for anger, fear and sad emotions. For MFCC, the lowest accuracy is for disgust emotion i.e. 60% and the rest is detected as fear emotion with 26.7%, as neutral with 13.3%. Whereas for T-MFCC lowest accuracy is for neutral with 80% and the rest is recognized as anger, sad with 6.7% each. Similarly, for female speech, using MFCC highest accuracy of 93.3% is for anger, sad emotions and the lowest is 73.3% for fear. For T-MFCC, the highest accuracy of 93.3% is obtained for anger, disgust, sad emotions, and the lowest accuracy is 80% for fear emotion.

Table 3.3: Confusion matrix of the SER system using LPC and T-LPC (proposed) feature extraction techniques

Emotion	Gender	Classification accuracy (%)									
		Anger		Fear		Disgust		Neutral		Sad	
		LPC	T-LPC (Proposed)	LPC	T-LPC (Proposed)	LPC	T-LPC (Proposed)	LPC	T-LPC (Proposed)	LPC	T-LPC (Proposed)
Anger	Male	**80**	**93.3**	0	0	0	6.7	20	0	0	0
	Female	**80**	**93.3**	13.3	0	6.7	6.7	0	0	0	0
Fear	Male	26.6	6.7	**53.3**	**66.7**	6.7	13.3	13.3	13.3	0	0
	Female	6.7	6.7	**60**	**66.7**	26.6	20	0	0	6.7	0
Disgust	Male	0	0	6.7	0	**80**	**100**	0	0	13.3	0
	Female	0	0	6.7	20	**80**	**66.6**	13.3	6.7	0	6.7
Neutral	Male	0	0	0	0	0	0	**93.3**	**80**	6.7	20
	Female	0	0	6.7	0	13.3	6.7	**80**	**93.3**	0	0
Sad	Male	0	0	13.3	0	0	0	20	0	**66.7**	**100**
	Female	0	0	0	0	13.3	0	6.7	6.7	**80**	**93.3**

From table 3.3, it can be observed that the SER system, after applying the Teager energy operator, the accuracy in case of all the emotions is increased for both male and female speakers. The highest accuracy of 100% is achieved for disgust and sad emotions for male speakers using T-LPC. For anger emotion, 93.3% and for fear emotion, 66.7% accuracies are achieved for both male and female speakers using T-LPC. These accuracies are correspondingly higher compared to the baseline LPC based SER system.

From table 3.4, the classification accuracy of the emotion recognition system for male speech is 100% for anger and sad emotion using LPCC, whereas using T-LPCC 100% accuracy is achieved for anger, disgust and sad emotions. Similarly, for female speech 100% accuracy is achieved for disgust and neutral and lowest is 80% for fear emotion, whereas, using T-LPCC accuracy is almost high i.e. 93.3% for anger, disgust, neutral and also sad emotions. From this analysis, it can be inferred that by using T-LPCC, though the classification accuracy for any of the emotions is not 100%, the accuracy of most of the emotions is comparably high compared to LPCC feature extraction.

Table 3.4: Confusion matrix of the SER system using LPCC and T-LPCC (proposed) feature extraction techniques

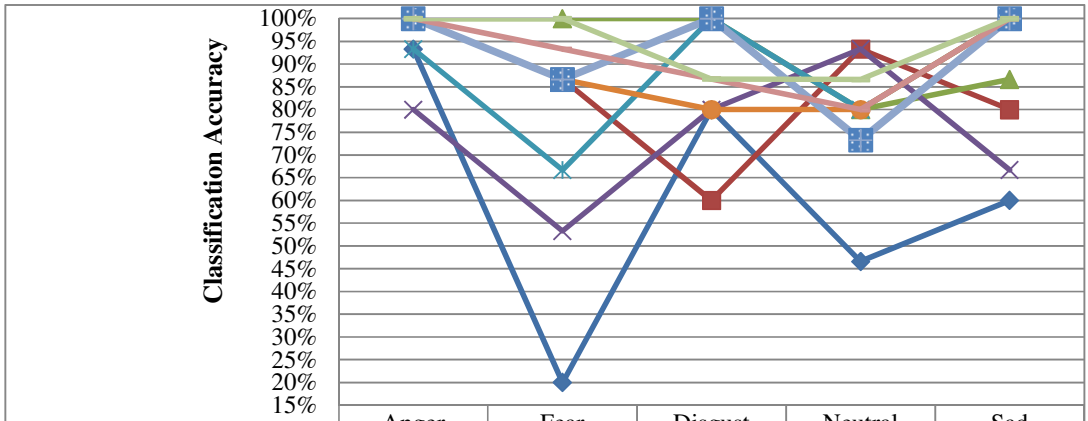
Classification accuracy (%)											
		Anger		Fear		Disgust		Neutral		Sad	
Emotion	Gender	LPCC	T-LPCC (Proposed)	LPCC	T-LPCC (Proposed)	LPCC	T-LPCC (Proposed)	LPCC	T-LPCC (Proposed)	LPCC	T-LPCC (Proposed)
Anger	Male	**100**	**100**	0	0	0	0	0	0	0	0
	Female	**86.7**	**93.3**	0	0	13.3	6.7	0	0	0	0
Fear	Male	6.7	6.7	**86.6**	**86.6**	6.7	6.7	0	0	0	0
	Female	0	6.7	**80**	**86.6**	13.3	0	6.7	0	0	6.7
Disgust	Male	20	0	0	0	**80**	**100**	0	0	0	0
	Female	0	0	0	6.7	**100**	**93.3**	0	0	0	0
Neutral	Male	13.3	20	0	0	0	0	**80**	**73.3**	6.7	6.7
	Female	0	6.7	0	0	0	0	**100**	**93.3**	0	0
Sad	Male	0	0	0	0	0	0	0	0	**100**	**100**
	Female	0	0	0	0	0	0	6.7	6.7	**93.3**	**93.3**

Table 3.5: Confusion matrix of the SER system using proposed MFCC-RASTA-PLP and T- MFCC-RASTA-PLP feature extraction techniques

Classification accuracy (%)											
		Anger		Fear		Disgust		Neutral		Sad	
Emotion	Gender	MFCC RASTA-PLP	T-MFCC-RASTA-PLP	MFCC RASTA-PLP	T-MFCC RASTA-PLP	MFCC RASTA-PLP	T-MFCC-RASTA-PLP	MFCC RASTA-PLP	T-MFCC RASTA-PLP	MFCC RASTA-PLP	T-MFCC RASTA-PLP
Anger	Male	**100**	**100**	0	0	0	0	0	0	0	0
	Female	**93.3**	**100**	0	0	6.7	0	0	0	0	0
Fear	Male	0	0	**93.3**	**100**	6.7	0	0	0	0	0
	Female	0	0	**93.3**	**100**	0	0	0	0	6.7	0
Disgust	Male	0	0	13.3	0	**86.7**	**86.7**	0	13.3	0	0
	Female	6.7	0	6.7	6.7	**86.6**	**93.3**	0	0	0	0
Neutral	Male	0	6.7	6.7	0	0	0	**80**	**86.6**	13.3	6.7
	Female	0	0	0	0	6.7	6.7	**93.3**	**93.3**	0	0
Sad	Male	0	0	0	0	0	0	0	0	**100**	**100**
	Female	0	0	0	0	0	0	6.7	6.7	**93.3**	**93.3**

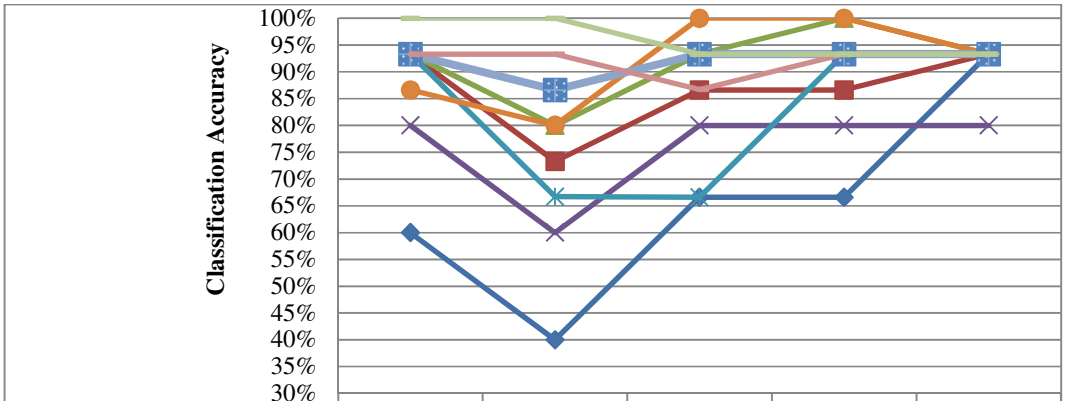
From table 3.5, the classification accuracy for male speech using MFCC-RASTA-PLP is 100% for anger, sad emotions and almost high as 93.3% for fear, whereas for female speech 93.3% accuracy is obtained for anger, fear, sad and the lowest accuracy is achieved for disgust with 86.6%. Similarly, by using T-MFCC-RASTA-PLP, further, the accuracy increases to 100% for anger, fear and sad emotions for male speech, and for female speech 100% accuracy for anger, fear and 93.3% for disgust, neutral, sad emotions.

Tables 3.1 to 3.5 shows the classification accuracies of the different emotions using the existing and the proposed feature extraction techniques for both male and female speech, these results can be concisely shown as in figures 3.7 and 3.8.



	Anger	Fear	Disgust	Neutral	Sad
Pitch	93.30%	20.00%	80.00%	46.60%	60.00%
MFCC	100.00%	86.70%	60.00%	93.30%	80.00%
T-MFCC	100.00%	100.00%	100.00%	80.00%	86.60%
LPC	80.00%	53.30%	80.00%	93.30%	66.70%
T-LPC	93.30%	66.70%	100.00%	80.00%	100.00%
LPCC	100.00%	86.60%	80.00%	80.00%	100.00%
T-LPCC	100.00%	86.60%	100.00%	73.30%	100.00%
MFCC-RASTA-PLP	100.00%	93.30%	86.70%	80.00%	100.00%
T-MFCC-RASTA-PLP	100.00%	100.00%	86.70%	86.60%	100.00%

Figure 3.7: Variation of SER accuracy for Pitch, MFCC, LPC, LPCC, MFCC-RASTA-PLP, T-MFCC, T-LPC, T-LPCC and T-MFCC-RASTA-PLP based SER system for different stressed emotions of male speakers



	Anger	Fear	Disgust	Neutral	Sad
Pitch	60.00%	40.00%	66.60%	66.60%	93.30%
MFCC	93.30%	73.30%	86.60%	86.60%	93.30%
T-MFCC	93.30%	80.00%	93.30%	100.00%	93.30%
LPC	80.00%	60.00%	80.00%	80.00%	80.00%
T-LPC	93.30%	66.70%	66.60%	93.30%	93.30%
LPCC	86.60%	80.00%	100.00%	100.00%	93.30%
T-LPCC	93.30%	86.60%	93.30%	93.30%	93.30%
MFCC-RASTA-PLP	93.30%	93.30%	86.70%	93.30%	93.30%
T-MFCC-RASTA-PLP	100.00%	100.00%	93.30%	93.30%	93.30%

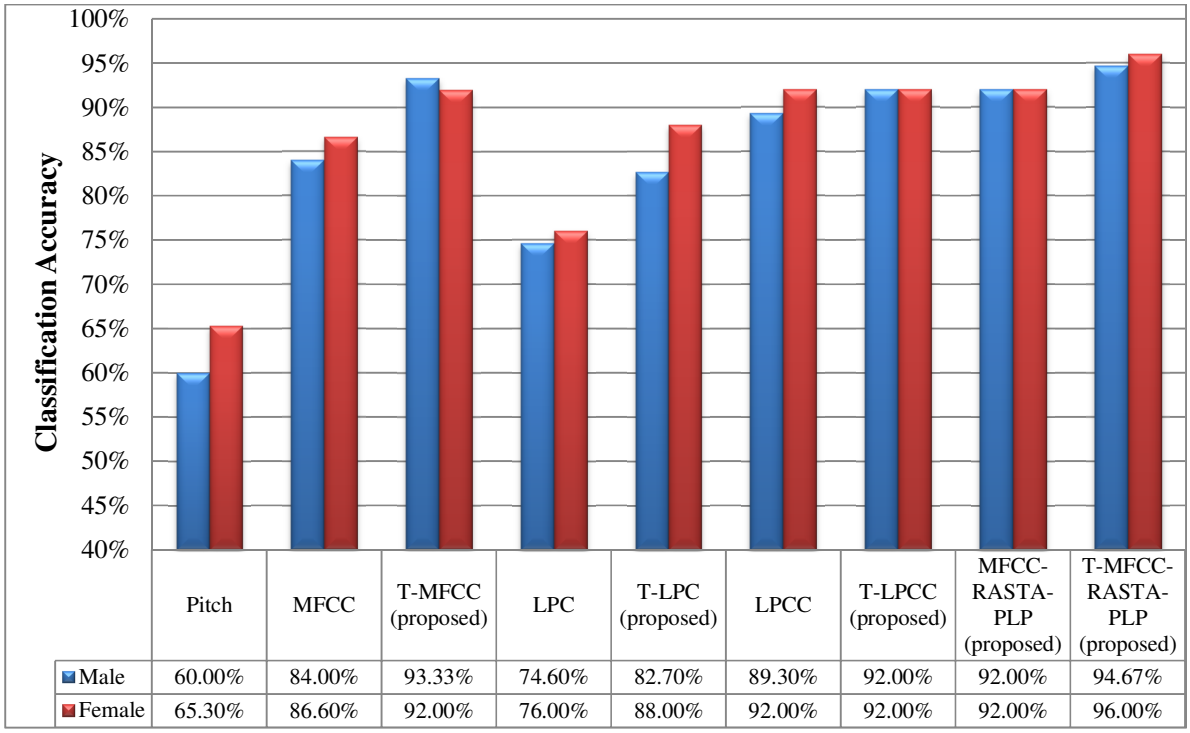
Figure 3.8: Variation of SER accuracy for Pitch, MFCC, LPC, LPCC, MFCC-RASTA-PLP, T-MFCC, T-LPC, T-LPCC and T-MFCC-RASTA-PLP based SER system for different stressed emotions of female speakers

From figure 3.7, for male speech, using pitch feature extraction technique, the anger emotion is detected with 93.3% accuracy, whereas fear with 20%, disgust with 80%, neutral with 46.6% and sad with 60% accuracy respectively. Using MFCC, anger emotion is recognized with 100%, fear with 86.7%, disgust with 60%, neutral with 93.3% and sad with 80% accuracy. By using LPCC, anger and sad emotions are identified with 100%, fear with 86.6% and disgust, neutral with 80% accuracy. When the MFCC-RASTA-PLP technique is used, 100% accuracy is obtained for anger, sad emotions, 93.3% for fear emotion, 86.67% for disgust and neutral with 80%. After combining these features with TEO, the accuracy of the system improves compared to the existing ones. T-MFCC provides an accuracy of 100% for anger, fear and disgust emotions, 80% accuracy for neutral speech and 86.6% for sad emotion. T-LPCC gives an accuracy of 100% for anger, disgust and sad emotions, 86.6% for

fear emotion and 73.3% for neutral speech. And T-MFCC-RASTA-PLP detects anger, fear, sad emotions with 100% accuracy, and with 86.7% accuracy for disgust emotion and neutral speech.

Similarly from figure 3.8, it can be observed that in the case of female speech, the emotion recognition system using pitch feature extraction technique, the anger emotion is detected with 60% accuracy, fear with 40%, disgust and neutral with 66.6% and sad emotion with 93.3% accuracy. Using MFCC, anger and disgust with 86.6%, fear with 73.3%, neutral speech and sad emotion with 93.3% accuracy. By using LPCC, anger emotion is identified with 86.6% accuracy, fear with 80%, disgust and neutral speech with 100%, and sad emotion with 93.3% accuracy. When the MFCC-RASTA-PLP technique is used, the anger, fear, neutral and sad emotions are detected with 93.3% accuracy and disgust emotion with 86.7% accuracy. T-MFCC provides an accuracy of 100% for anger emotion and neutral speech, 86.6% for fear and disgust, and 93.3% for sad emotion. T-LPCC gives an accuracy of 93.3% for anger, disgust, neutral and sad emotions, 86.6% for fear emotion. And the T-MFCC-RASTA-PLP based emotion recognition system detects anger and fear emotions with 100% accuracy, whereas disgust, neutral and sad emotions with 93.3% accuracy.

The overall accuracy of the proposed speech emotion recognition system with the comparison of all the feature extraction techniques is depicted in figure 3.9. The classification accuracy of the recognition system using the pitch feature extraction technique is 60% for male and 65.3% for female speech. This accuracy is improved as 84% in the case of male, 86.6% female speech by using MFCC and further improved by using T-MFCC with 93.3% for both male and female speech data. Using LPC the SER accuracy is 74.6% for male, 76% for female and using T-LPC, the accuracy is 82.7% for male, 88% for female respectively. In the case of LPCC, the accuracy is 89.3% for male, 92% for female speech and by using T-LPCC 92% accuracy is obtained for both male and female speech. The improvement is in the case of male speech using T-LPCC compared to LPCC. The MFCC-RASTA-PLP based system provided an accuracy of 93.3% for male, 92% for female speech and using T- MFCC-RASTA-PLP with 96% and 93.3% accuracy respectively.



	Pitch	MFCC	T-MFCC (proposed)	LPC	T-LPC (proposed)	LPCC	T-LPCC (proposed)	MFCC-RASTA-PLP (proposed)	T-MFCC-RASTA-PLP (proposed)
Male	60.00%	84.00%	93.33%	74.60%	82.70%	89.30%	92.00%	92.00%	94.67%
Female	65.30%	86.60%	92.00%	76.00%	88.00%	92.00%	92.00%	92.00%	96.00%

Figure 3.9: Comparison of the proposed SER system using T-MFCC, T-LPC, T-LPCC, MFCC-RASTA-PLP, T-MFCC-RASTA-PLP, Pitch, MFCC, LPC and LPCC

3.4 Summary:

A spectral feature fusion of MFCC and RASTA-PLP as MFCC-RASTA-PLP, and a combination of spectral features (MFCC, LPCC, MFCC-RASTA-PLP) and TEO as T-MFCC, T-LPCC and T-MFCC-RASTA-PLP are proposed with GMM classifier for the detection of stressed emotions (anger, fear, disgust and sadness). The Teager Energy operator is mainly designed to detect the stressed emotions and hence, it provides improved accuracy compared to the existing techniques. For male and female speakers, respectively, the performance analysis using the EMO-DB database showed an accuracy of 92% & 92% for MFCC-RASTA-PLP and 94.67% & 96% using T-MFCC-RASTA-PLP based stressed emotion recognition systems. The accuracy of the SER system using the T-MFCC feature is 93.3% &

92% and using the T-LPCC feature is 92% for both male and female speakers. These accuracies are comparatively higher than that of the existing feature extraction techniques such as Pitch (60% & 65.3%), MFCC (84% & 86.6%) and LPCC (89.3% & 92%) based SER systems. For evaluating the proposed system, only the stressed emotions of the EMO-DB database have been considered. Whereas, to develop a sophisticated SER system the entire set of emotions has to be considered for system analysis. Therefore, in the next chapter a speech emotion recognition system is developed by considering more number of emotions. Further, the feature fusion is carried out in SER analysis to acquire higher classification accuracies.

Chapter 4

Speech Emotion Recognition using Semi-NMF Feature Optimization

In this chapter, a speech emotion recognition system is proposed using semi-nonnegative matrix factorization (semi-NMF) with singular value decomposition (SVD) initialization to optimize the speech features. The speech features considered for the proposed SER system are Mel-frequency cepstral coefficients, linear prediction cepstral coefficients, and Teager energy operator-autocorrelation (TEO-AutoCorr). The k-nearest neighborhood (K-NN) and support vector machine (SVM) pattern recognition algorithms with supervised learning are used for the classification of emotions with a 5-fold cross-validation scheme. SER system developed using the semi-NMF algorithm is validated in terms of classification accuracy and number of speech features.

4.1 Motivation

The SER systems with single feature sets provide low accuracies rather than the system with multiple feature fusion. Thus the combination of the feature fusion for SER has become a common practice. Even though the classification accuracy of the SER system increases due to feature fusion, the computational overhead also increases on the classifier. The reason is - only some of the features are useful for SER analysis whereas many other features have no role in emotion recognition. This leads to the curse of dimensionality and a decrease in the SER performance, where the performance of the system in terms of classification accuracy is decreased after a particular feature dimension threshold with an increase in the dimension of the feature set. Therefore, it is always preferable to perform feature selection or optimization of the feature sets before classifying emotions.

Feature dimension reduction is the best way to solve the problem of high dimensionality, but the reduction of the number of feature vectors causes an uncertain loss in the information and subsequently leads to instability in the performance of the system. This problem can be overcome by using linear transformation techniques. If an n-dimensional input feature vector $x = [F_1, F_2, \ldots, F_n]^T$ is considered, then the transformed output vector will be $y = [b_1, b_2, \ldots, b_r]^T$, where $r \ll n$ and 'r' is the reduced dimension. For this purpose, many optimization techniques are developed in machine learning to acquire the most optimal feature sets that improve SER accuracy. Therefore, an SER system with Semi-NMF feature optimization is proposed for feature dimension reduction.

4.2 Proposed SER System using Semi-Non Negative Matrix Factorization (Semi-NMF)

A conventional SER system consists of only three stages: speech preprocessing, feature extraction, and classification [8], [23]. Most of the existing SER systems use the combined set of speech features for emotion recognition. This increases the computational

overhead on the classification model. To overcome this drawback, the semi-NMF optimization technique is incorporated in the development of the SER system before classifying the features for obtaining the emotions, as shown in Figure 4.1. In the proposed system, after feature extraction before classification of emotions, a semi-NMF with SVD initialization feature optimization algorithm is used to reduce the initial huge feature sets.

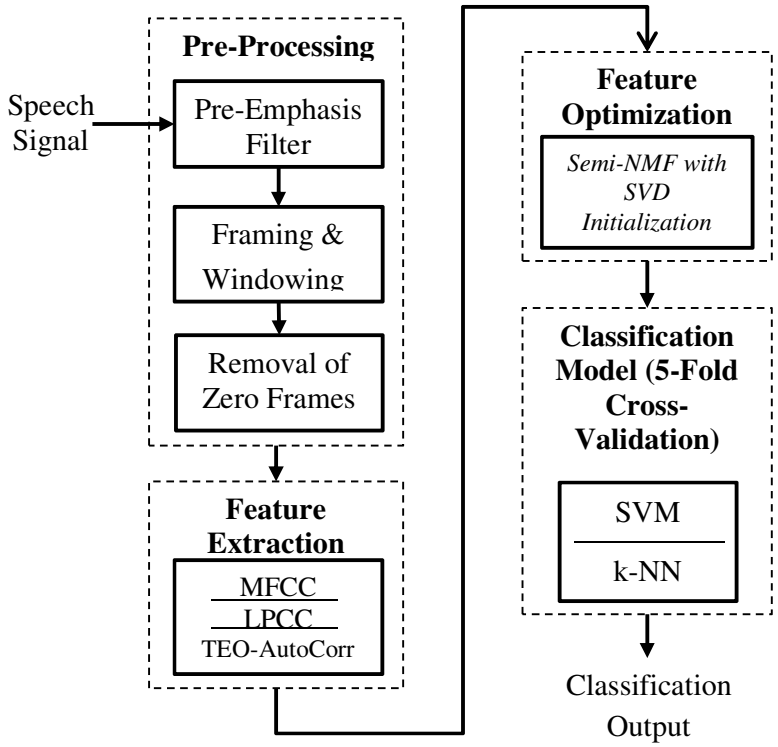


Figure 4.1: Proposed Speech Emotion Recognition System using Semi-NMF

4.2.1 Feature Extraction

In SER, the emotional relevant speech features are extracted from speech signals using the feature extraction process. To obtain the emotional contents from a speech signal, a particular set of features can be extracted by applying various signal processing techniques. 24 MFCC, 21 LPCC, and 20 TEO-AutoCorr features are extracted in the development of the proposed SER system.

MFCC features contribute mostly to SER system development, as these features are designed based on the human ear speech perception. In MFCCs, initially, the speech signal frames are transformed into the frequency domain using DFT and the transformed frames are fed to the Mel-filter bank to convert the log frequency-scale to the Mel-frequency scale, which mimics the perception of a human ear [23]:

$$mel(f) = 2595 \log_{10}(1 + \frac{f}{700}) \quad \text{or} \quad mel(f) = 1127 \ln(1 + \frac{f}{700}) \qquad (4.1)$$

Here, f is the frequency of the transformed speech signal. These transformed Mel-frequency domain features are further converted to the cepstrum domain using the discrete cosine transform (DCT). A total of 12 MFCC features are extracted in this process. It is well known that the difference between the consecutive MFCC features, which are termed as Δ (delta) features, contributes to efficient emotion recognition. Hence, a total of 24 features with 12 MFCC and 12 Δ (delta) features are extracted.

In the extraction of LPCCs, initially, linear predictive analysis is performed on the speech signal. The basic idea behind the linear predictive analysis is that the nth speech sample can be estimated by a linear combination of its previous p samples as shown in the following equation:

$$x[n] \approx a_1 x[n-1] + a_2 x[n-2] + a_3 x[n-3] + \cdots + a_p x[n-p] \qquad (4.2)$$

The LPCC feature extraction is designed to obtain '21' features and is used in this analysis.

4.2.1.1 Teager Energy Operator Auto-Correlation Features (TEO-AutoCorr):

Even though MFCCs and LPCCs are widely used for SER, a few of the stressed emotions like anger or anxiety could not be analyzed properly. Therefore, TEO features are also used in this work. The TEO-AutoCorr feature extraction is shown in Figure 4.2.

Teager proposed an energy operator i.e., a measure of speech signal energy based on his experiments known as the Teager Energy Operator [46], [47]. The energy operator is [47], [49],

$$\psi(x(t)) = \left(\frac{d}{dt}x(t)\right)^2 - x(t)\left(\frac{d^2}{dt^2}x(t)\right) \quad \text{(or)} \quad \psi(x[n]) = x^2[n] - x[n]x[n+1] \tag{4.5}$$

Here, '$x(t)$' and '$x[n]$' are the speech signals in the continuous & discrete domain.

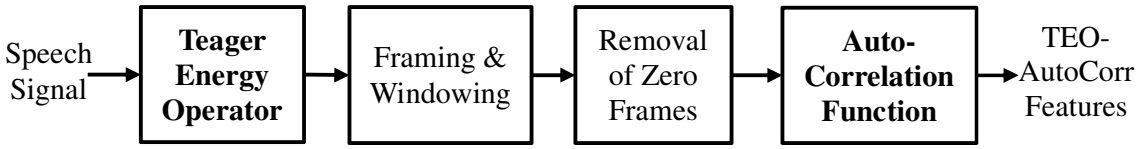


Figure 4.2: TEO-AutoCorr Feature Extraction

The auto-correlation function is the correlation of a signal with a delayed copy of itself as a function of delay. Informally, it is the similarity between observations as a function of the time lag between them and is given by:

$$R_{xx}(k) = \sum_{n=k}^{M-1} \mathrm{s[n]s[n-k]} \tag{4.6}$$

Here, '$s[n]$' is the input signal to the function and $'k'$ is the delay parameter. When the frames of the Teager energized signal are given to the autocorrelation function, the correlation between the adjacent frames is obtained. If the correlation is high, the energy of the speech signal is further increased, resulting in the TEO-AutoCorr features. All the extracted features say $[F_1, F_2, \ldots\ldots, F_k]$, are further fed to the feature optimization block.

4.2.2 Feature Optimization using Semi-NMF with SVD Initialization:

Non Negative Matrix Factorization (NMF) is one of the well-known feature optimization technique in machine learning for data dimension reduction. But the NMF algorithm accepts only the non-negative data as input. Whereas, the features (MFCC, LPCC & TEO-AutoCorr) contains mixed signs (both positive and negative) and NMF could not be used as such. The semi-NMF technique is a variant of the NMF algorithm and can be used for speech feature optimization. In the proposed SER system, the semi-NMF algorithm using SVD initialization is employed to optimize the speech features. Semi-NMF has been widely used in many data processing applications like data analysis and clustering [122]. The data matrix, i.e. a feature matrix $M = (F_1, F_2, \ldots\ldots, F_k)$ with k as the feature vectors that are unconstrained (i.e. it may have mixed signs), is considered. A factorization that is referred to as semi-NMF in [122], in which V is restricted to be nonnegative while not restrict the signs of U, is proposed.

Semi-NMF can be defined as follows: Given a matrix $M \in \mathbb{R}^{m\times k}$ and a factorization rank r, solve

$$\min_{U\in\mathbb{R}^{m\times r}, V\in\mathbb{R}^{r\times k}} ||M - UV||_F^2 \text{ such that } V \geq 0 \quad (4.7)$$

where $||.||_F$ is the Frobenius norm and $V \geq 0$ means that V is component-wise non-negative. The concept of Semi-NMF is motivated from the perspective of k-means clustering

that can be applied to an input feature vector 'M' to obtain cluster centroids, $U = \{u_1, u_2, \dots\dots, u_r\}$ and 'V' be the cluster indicators [122],

$$V = \begin{cases} 1 & if\ F_i \in cluster\ u_r \\ 0 & otherwise \end{cases} \tag{4.8}$$

However, there are convergence issues in this method, due to which a different initialization technique rather than k-means can be used since the initialization of U and V matrices are important to obtain an optimal solution for the factorization problem. The semi-nonnegative rank of matrix M can be denoted by $M = UV$ with $U \in \mathbb{R}^{m \times r}$, $V \in \mathbb{R}^{r \times k}$ and $V \geq 0$. To summarize,

$$NMF \rightarrow X_+ \approx U_+ V_+^T \tag{4.9}$$

$$Semi - NMF \rightarrow X_\pm \approx U_\pm V_+^T \tag{4.10}$$

In other words, NMF has both U and V with nonnegative values, whereas semi-NMF has U consisting of both positive and negative values without any restriction and V with only nonnegative values. Accordingly, a singular value decomposition (SVD) and linear programming-based method are proposed to overcome the drawbacks of the basic semi-NMF for finding the optimal solution [123]. The semi-NMF technique with SVD initialization is discussed in Algorithm 4.1.

Algorithm 4.1: Semi-NMF using SVD Initialization

Input: A matrix $M \in \mathbb{R}^{m \times k}$, a factorization rank r.

Output: A rank- r Semi-NMF (U, V) of $M \approx UV$ with $V \geq 0$.

1: $[A, S, B^T] = svds(M, r)$; '$svds$' is a MATLAB function

2: For each $1 \leq i \leq r$: multiply $B(i,:)$ by -1 if $\min_j B(i,j) \leq \min_j \left(-B(i,j)\right)$;

3: Let (y^*, ε^*) be the optimal solution of the following optimization problem

$\min_{y \in \mathbb{R}^r, \varepsilon \in \mathbb{R}_+} \varepsilon$ such that $(B(:,j) + \varepsilon e)^T y \geq 1 \; \forall \, j$ such that $B(:,j) + \varepsilon e \neq 0$;

% if $\epsilon^* = 0$ ($\Leftrightarrow$ B is semi-nonnegative) then the heuristic is optimal

4: $x = (B + \varepsilon^* 1_{r \times k})^T y^* \geq 1$; % $1_{r \times k}$ is the $r - by - k$ matrix of all ones

5: $\propto_i = \max\left(0, max_j \frac{-B(i,j)}{x(j)}\right)$ for all $1 \leq i \leq r$;

6: $V = B + \propto x^T$;

7: $U \leftarrow \text{argmin}_{X \epsilon \mathbb{R}^{m \times r}} ||M - XV||_F^2$

In step 1, we apply SVD on the data or feature matrix to obtain the left singular matrix (A), diagonal matrix (S), and right singular matrix (B) considering rank-r approximation. Among these, matrix B is considered for further analysis.

In step 2, the rows of matrix B are flipped. Further, in step 3, the actual optimization takes place, i.e. a heuristic for finding the optimal solution (y^*, ε^*) with r,

$$\min_{y\in\mathbb{R}^r \varepsilon\in\mathbb{R}_+} \varepsilon \text{ such that } (B(:,j)+\varepsilon e)^T y \geq 1 \;\forall\, j \text{ such that } B(:,j)+\varepsilon e \neq 0; \tag{4.11}$$

Here e is the vector of all ones. If the value of ε^* is too small the probability of B being a semi-nonnegative matrix is high. Eq. (4.11) is solved using a bisection method on the variable ε^*. In this method, if $\varepsilon^*= 0$ initially, then an optimal semi-NMF can be obtained. Once the optimal solution (y^*, ε^*) is obtained, the matrices V and further the desired optimal solution U can be obtained in the consecutive steps.

In this work, the rank-r of the semi-NMF is chosen to acquire optimum performance. The MFCC, LPCC, and TEO-AutoCorr features will be scaled down to r number of features each. Using semi-NMF, the $m \times 24$ feature matrix is factorized into U_{mel} and V_{mel} matrices with $m \times r_1$ and $r_1 \times 24$. Likewise, the LPCC feature vector with $m \times 21$ dimensions, using semi-NMF, is factorized into U_{lp} and V_{lp}feature and coefficient matrices with $m \times r_2$ and $r_2 \times 21$. Similarly, in the case of TEO-AutoCorr features, using semi-NMF, the $m \times 20$ feature vector matrix is factorized into U_{teo} and V_{teo} feature and coefficient matrices with $m \times r_3$ and $r_3 \times 20$. Here, the matrices 'U_{mel}', 'U_{lp}' and U_{teo} are the desired optimal MFCC, LPCC, and TEO-AutoCorr feature vectors consisting of both the positive and negative data. The ranks r_1, r_2 and $r_3$are chosen based on the type of the classifier (SVM or k-NN) and the feature set chosen (i.e. MFCC or LPCC or TEO-AutoCorr) by validating the performance of the SER system.

4.2.3 Classification

Many classification techniques are proposed in machine learning for pattern recognition [25], [102]. k-nearest neighborhood (K-NN) and support vector machine (SVM) classification techniques are used in the proposed system to classify the emotions after feature optimization. Supervised learning is used in these classifiers i.e., specifying the label of the emotion for the training process. To ensure that, there is no overfitting problem in the proposed model, k-fold cross-validation is used during the classification process.

4.2.3.1 k-Nearest Neighborhood

K-Nearest Neighbour algorithm is a simple supervised machine learning technique used for classification. In this algorithm when a new data or test case is to be predicted, the resemblance between the test case and the training data is carried out. The category of the test case is assigned with one of the existing categories that have the best similarity. In the K-NN algorithm, initially, data is trained with correspondence to the labels of different categories. Whenever a new test input arrives, it is assigned to any of the available categories based on the similarity index. K-NN can also be used to solve regression problems but mostly used for classification. During the training phase, K-NN only stores the data and do not perform any learning. Only during classification, K-NN categorizes the data and due to this, it is also called a lazy learner algorithm.

The similarity between the test data and training data is performed using distance measure algorithms. The different distance measure techniques are Euclidean, Manhattan, Minkowski, etc. Among these, the Euclidean distance measure technique is the one that is mostly used for the K-NN task. This is represented as:

$$\sqrt[2]{\sum_{i=1}^{k}(|x_i - y_i|)^2} \tag{4.12}$$

Here, x_i and y_i are the data points of the test data and the nearest neighbor of training data, and k is the number of nearest neighbors chosen.

Let us consider a classification problem with two categories A and B. If there is a new test data point, using K-NN, this test data point can be assigned to any one of these categories.

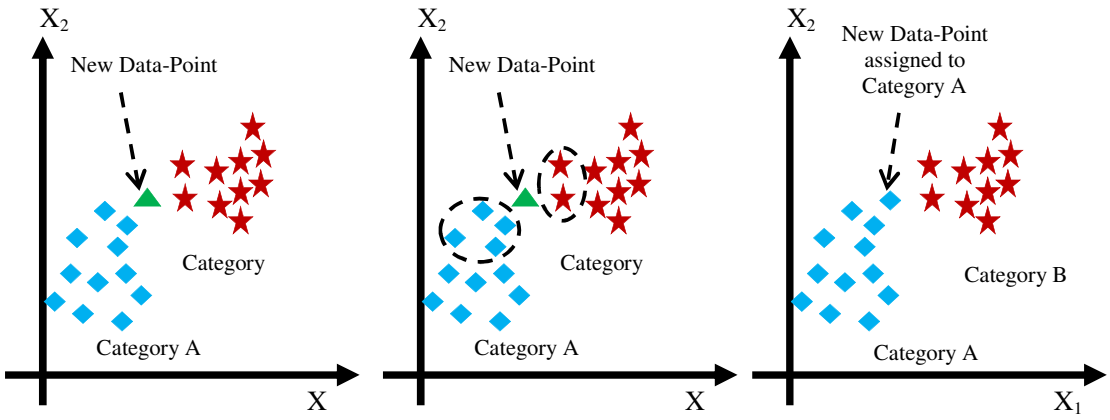


Figure 4.3: K-NN illustration

Figure 4.3 shows the illustration of the K-NN technique. Here, consider k=6 nearest neighbors to select the nearest neighbors of the new data point. The four data points from category A and two data points from category B are nearest to the new data point. According to the principle of K-NN, the category with the highest number of nearest neighbors is assigned to the new data point i.e., category A.

Similarly in SER, the speech features are data points i.e., X_1, X_2 are the features and the categories are emotions of the database considered. The K-NN illustration is shown only in two-dimension space. For SER, multiple speech features in multi-dimension space using the K-NN principle is carried out. For the proposed work, k= 6 nearest neighbors are chosen to carry out the emotion classification using K-NN

4.2.3.2 Support Vector Machine

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning.

The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane. SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called support vectors, and hence algorithm is termed a Support Vector Machine. Consider the below diagram in which two different categories are classified using a decision boundary or hyperplane:

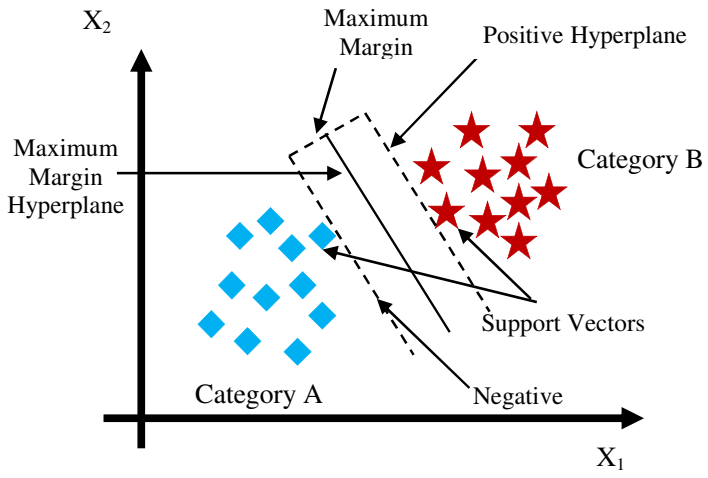


Figure 4.4: SVM illustration

SVM algorithm can be used for Face detection, image classification, text categorization, etc. SVM can be of two types:

i. **Linear SVM:** Linear SVM is used for linearly separable data, which means if a dataset can be classified into two classes by using a single straight line, then such data is termed as linearly separable data, and classifier is used called as Linear SVM classifier.
ii. **Non-linear SVM:** Non-Linear SVM is used for non-linearly separated data, which means if a dataset cannot be classified by using a straight line, then such data is termed as non-linear data and the classifier used is called a Non-linear SVM classifier.

There can be multiple lines/decision boundaries to segregate the classes in n-dimensional space, but we need to find out the best decision boundary that helps to classify the data points. This best boundary is known as the hyperplane of SVM. The dimensions of the hyperplane depend on the features present in the dataset, which means if there are 2 features as shown in figure 4.4, then the hyperplane will be a straight line. And if there are 3 features, then the hyperplane will be a 2-dimension plane.

The hyperplane is created such that has a maximum margin, which means the maximum distance between the data points. The data points or vectors that are the closest to the hyperplane and which affect the position of the hyperplane are termed as Support Vector. Since these vectors support the hyperplane, hence called a Support vector. In this work, a non-linear Gaussian kernel is chosen for the SVM classifier.

4.2.3.3 k-Fold Cross-Validation:

Cross-validation is a resampling procedure used to evaluate machine learning models on a limited data sample. This approach involves randomly dividing the set of observations into k groups, or folds, of approximately equal size. The first fold is treated as a validation set, and the method is fit on the remaining $'k-1'$ folds. Cross-validation is primarily used in applied machine learning to estimate the skill of a machine learning model on unseen data. That is, to use a limited sample to estimate how the model is expected to perform in general when used to make predictions on data not used during the training of the model. It is a popular method because it is simple to understand and because it generally results in a less biased or less optimistic estimate of the model skill than other methods, such as a simple train/test split. The 5-fold cross-validation schema is used to train the classifiers, in which the training and testing are carried out in 5-folds.

4.3 Simulation Results and Performance Evaluation

The EMO-DB and IEMOCAP datasets are considered in this work to analyze the proposed SER system performance. As discussed in Section 2, the 24 MFCC, 21 LPCC, and 20 TEO-AutoCorr features are extracted. The simulation parameters used in the development of the proposed speech emotion recognition system, i.e. for speech preprocessing, feature extraction, optimization, and classification, are shown in Table 4.1.

Initially, the 24 MFCC, 21 LPCC, and 20 TEO-AutoCorr features are optimized using the semi-NMF algorithm using SVD initialization, and the k-NN and SVM classifiers are used for classification of emotions. The classification accuracy and the number of speech features are used as the performance metrics to validate the proposed system. All the simulations are carried out on a computer with Intel Xeon CPU E3-1220 v3 of a 3.10 GHz 64-bit processor with 16 GB RAM.

Table 4.1: Simulation Parameters of Proposed SER System using Semi-NMF

Parameters	**Specifications**
Pre-Emphasis Filter	Coefficient, a = 0.97
Frame Size/ Length	256 samples
Frame Overlap	80 samples
Type of Window	Hamming
Mel-Filter Banks	20
Semi-NMF	SVD Initialization
Validation	5-Fold Cross-Validation
k-NN	k-Folds = 6 Distance Measure = Euclidean
SVM	Kernel = Gaussian

The corresponding results are shown in figures 4.5 to 4.8 and Tables 4.2 and 4.3. For the classification of emotions, 5- fold cross-validation scheme is adapted. Accordingly, the entire data is divided randomly into 5 folds, where the first fold is hold-out for validation and the rest 4 folds are to train the classifier. This process is repeated in 5 folds, i.e., 5 times, until the entire dataset is completely trained. In this work, the 5-fold cross-validation schema is used to train and test the accuracy of the proposed SER system. The evaluated score (i.e. the classification accuracy) at each fold is retained and, finally, the mean of these scores is calculated to obtain the overall classification accuracy of the proposed system.

The number of features into which the features get optimized depends on the rank of the semi-NMF. The choice of choosing the rank of the semi-NMF algorithm for optimizing the features to achieve high performance is very important. To decide the optimal rank, the optimization is performed individually on MFCC, LPCC, and TEO-AutoCorr features using different ranks of semi-NMF and these optimized features are classified using the classification models.

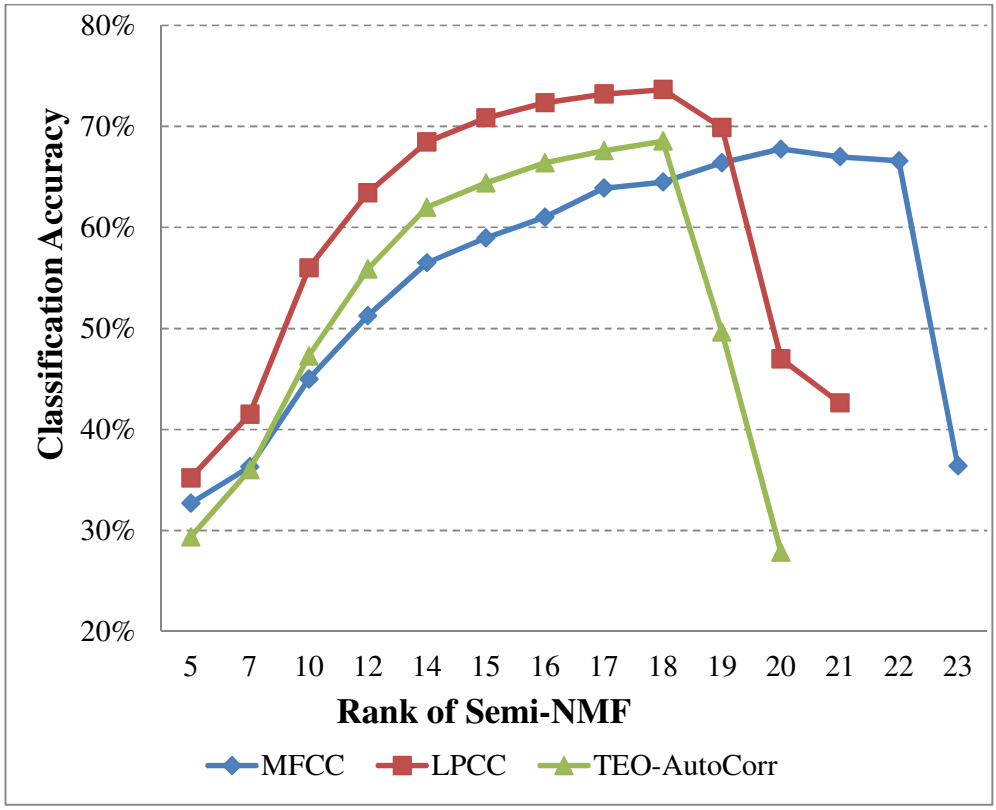


Figure 4.5: Variation of classification accuracy with Rank of Semi-NMF in the proposed SER system for MFCC, LPCC & TEO-AutoCorr Features using SVM for EMO-DB

Figures 4.5 to 4.8 show the deviation of emotion recognition accuracy for EMO-DB and IEMOCAP databases with different ranks of semi-NMF with which the MFCC, LPCC, and TEOAutoCorr features are optimized using SVM and k-NN classifiers in the proposed system. From these results, it is clearly understood that the performance of the SER system not only varies with the type of database used but also the classification model considered. From these figures, it can also be observed that the SER classification accuracy is increased with the rank of the semi-NMF considered, i.e. with an increase in the number of features, whereas after a particular rank of semi-NMF, the classification accuracy of the SER system is decreased, thus implying the curse of dimensionality. The rank at which utmost accuracy is obtained is considered to be the optimal rank.

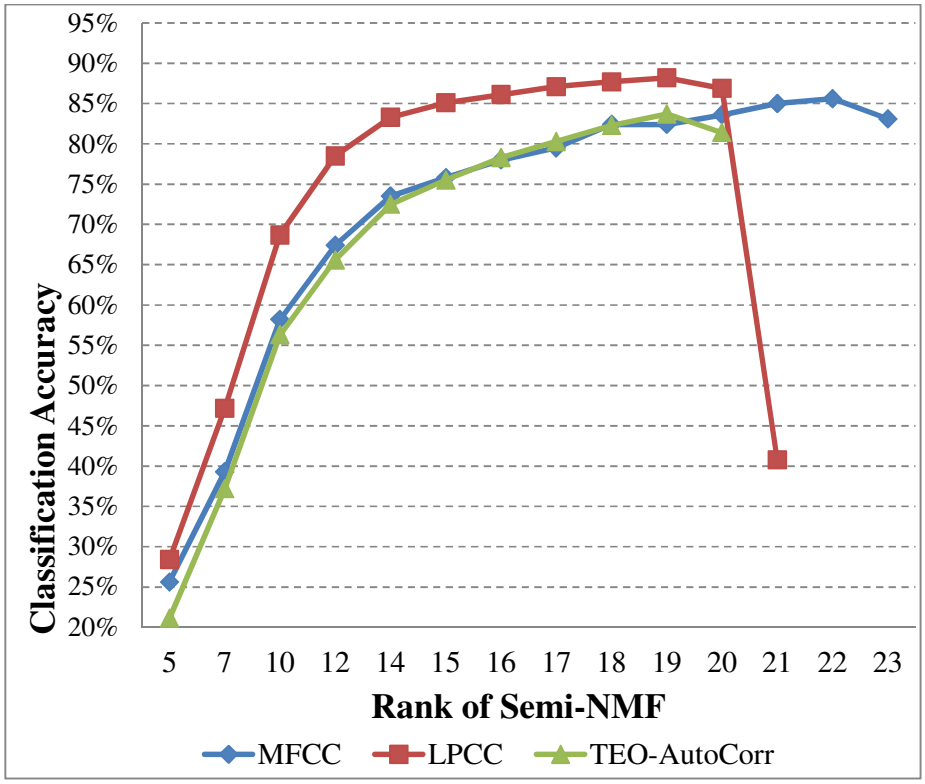


Figure 4.6: Variation of classification accuracy with Rank of Semi-NMF in the proposed SER system for MFCC, LPCC & TEO-AutoCorr Features using k-NN for EMO-DB

From Figures 4.5 and 4.6, for the EMO-DB database, the highest accuracy is achieved for MFCCs, when optimized with Rank-20 and Rank-22 for SVM and k-NN classifiers, as 67.76% and 85.6%, respectively. In the case of optimized LPCC and TEO-AutoCorr features using the SVM classifier, the highest accuracy is achieved with Rank-18 as 73.65% and 68.56%, respectively. Using the k-NN classifier, 88.2% and 83.7% classification accuracies are obtained for LPCC and TEO-AutoCorr features optimized at Rank-19. Therefore, the optimal ranks for MFCCs are 20 and 22 for SVM and k-NN, respectively. Similarly, for LPCCs and TEO, the optimal ranks are 19 and 18 using SVM and k-NN.

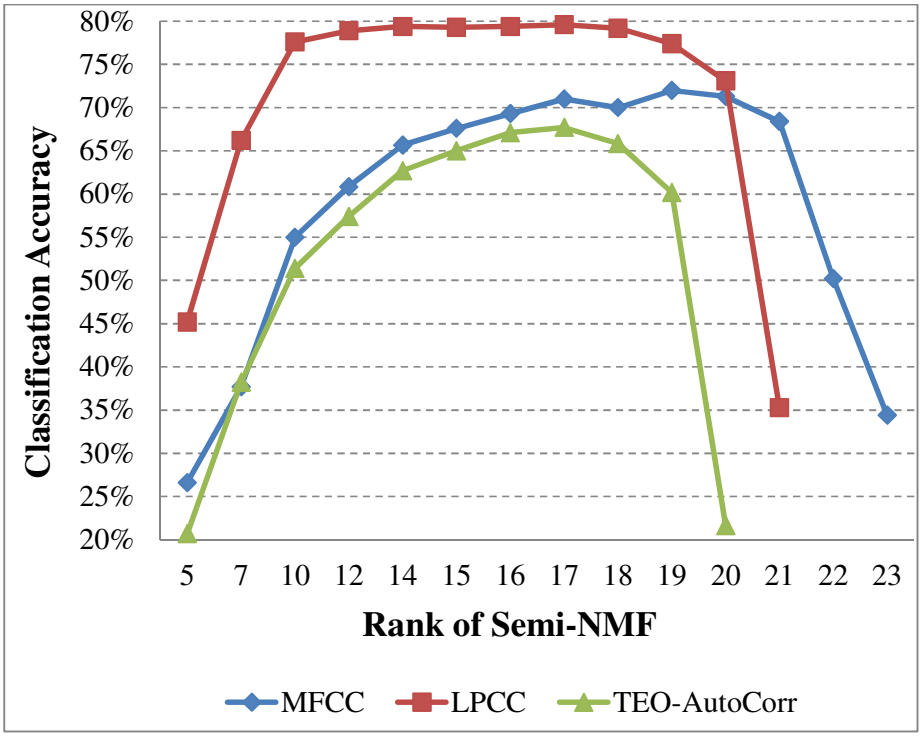


Figure 4.7: Variation of classification accuracy with Rank of Semi-NMF in the proposed SER system for MFCC, LPCC & TEO-AutoCorr Features using SVM for IEMOCAP

From Figures 4.7 and 4.8, for the IEMOCAP database, the highest accuracy is achieved for MFCCs, when optimized with Rank-19, as 72% and 74.1% for SVM and k-NN classifiers, respectively. In the case of optimized LPCCs, the highest accuracy is achieved with Rank-17 as 79.6% for the SVM classifier and with Rank-12 as 75% for the k-NN classifier. Likewise, for TEO-AutoCorr features, the highest accuracy is achieved with Rank-17 as 67.7% using the SVM classifier and with Rank-19 as 71% using the k-NN classifier.

The optimal rank for MFCC is 19 for both SVM and k-NN. Likewise, using SVM and k-NN classifiers, for LPCCs the optimal ranks are 17 and 12, whereas in the case of TEO features the optimal ranks are 17 and 19.

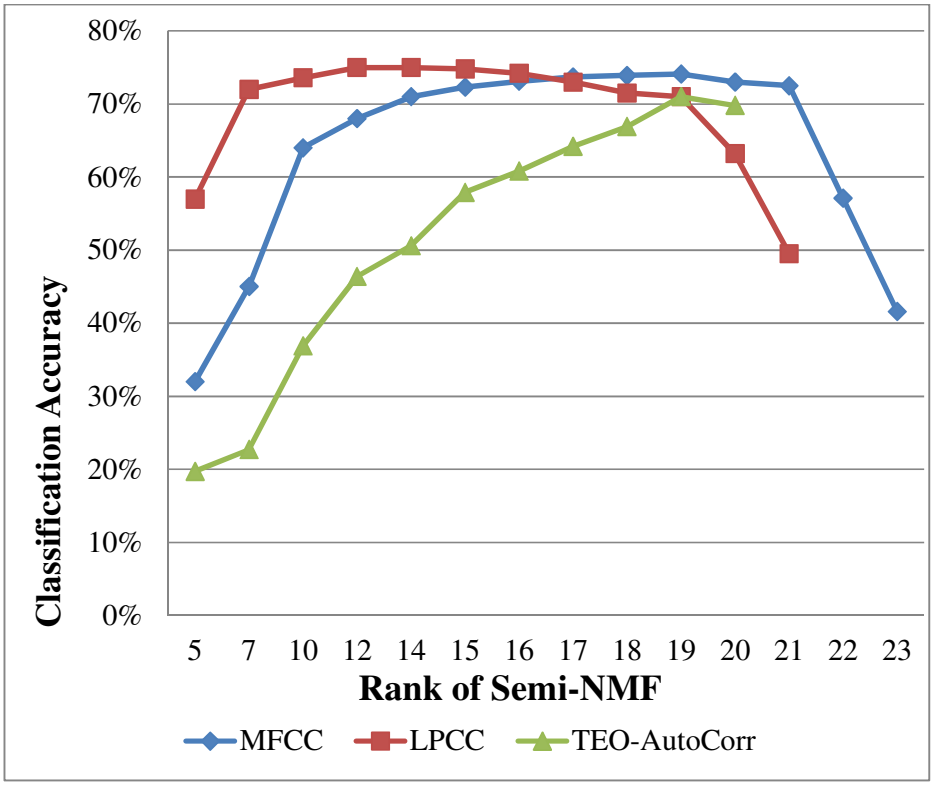


Figure 4.8: Variation of classification accuracy with Rank of Semi-NMF in the proposed SER system for MFCC, LPCC & TEO-AutoCorr Features using k-NN for IEMOCAP

Tables 4.2 and 4.3 show the results of the performance comparison of MFCCs, LPCCs, TEO-AutoCorr, and their combinations without optimization and with the semi-NMF optimization technique validated using SVM and k-NN classifiers. From these results, it is evident that by combining the features the performance is improved. This is the reason behind the extensive usage of huge feature sets for SER development.

Table 4.2: Comparison of Baseline and Proposed SER system with Semi-NMF for EMO-DB Database using SVM & K-NN Classifiers for Different Feature Sets

Optimization Techniques	**Features**	**SVM**		**k-NN**	
		No. of Features	Classification Accuracy	No. of Features	Classification Accuracy
Baseline	MFCC	24	47%	24	53%
	LPCC	21	45%	21	40.8%
	TEO-AutoCorr (TEO)	20	31.8%	20	27.3%
	MFCC+LPCC	45	64.2%	45	70.5%
	MFCC+TEO	44	62.5%	44	74.6%
	LPCC+TEO	42	46.47%	42	51.7%
	MFCC+LPCC+TEO	65	55.36%	65	69.9%
Semi-NMF with SVD	MFCC	20	67.76%	22	85.6%
	LPCC	**18**	**73.65%**	**19**	**88.2%**
	TEO-AutoCorr	18	68.56%	19	83.7%
	MFCC+LPCC	38	85%	41	89%
	MFCC+TEO	38	81.54%	41	87.8%
	LPCC+TEO	36	84.13%	38	88.7%
	MFCC+LPCC+TEO	**56**	**90.12%**	**60**	**89.3%**

From Table 4.2 for the EMO-DB database, it is observed that the highest accuracy is achieved with the feature fusion of the optimized MFCC, LPCC, and TEO-AutoCorr features with 56 features obtaining 90.12% accuracy using SVM and 89.3% accuracy using the k-NN classifier with 60 features. The minimum number of features at which the highest accuracy is obtained for the proposed system is 73.65% for SVM and 88.2% for k-NN, with LPCC features optimized at Rank-18 and Rank-19, respectively.

Table 4.3: Comparison of Baseline and Proposed SER system with Semi-NMF for IEMOCAP Database using SVM & K-NN Classifiers for Different Feature Sets

Optimization Techniques	**Features**	**SVM**		**k-NN**	
		No. of Features	Classification Accuracy	No. of Features	Classification Accuracy
Baseline	MFCC	24	44%	24	50.57%
	LPCC	21	41.1%	21	39.95%
	TEO-AutoCorr (TEO)	20	31.2%	20	29.2%
	MFCC+LPCC	45	45.2%	45	52.36%
	MFCC+TEO	44	43.4%	44	47.24%
	LPCC+TEO	42	42.3%	42	35.4%
	MFCC+LPCC+TEO	65	50.34%	65	55.63%
Semi-NMF with SVD	MFCC	19	72%	19	74.1%
	LPCC	**17**	**79.6%**	**12**	**75%**
	TEO-AutoCorr	17	67.7%	19	71%
	MFCC+LPCC	36	82.88%	31	74.98%
	MFCC+TEO	36	79.23%	38	75%
	LPCC+TEO	34	82.6%	31	74.32%
	MFCC+LPCC+TEO	**53**	**83.2%**	**60**	**78%**

Similarly, from Table 4.3 for the IEMOCAP database, the highest accuracy is achieved with the feature fusion of the optimized MFCC, LPCC, and TEO-AutoCorr features with 53 features obtaining 83.2% accuracy using SVM and 78% accuracy using the k-NN classifier with 50 features. The minimum number of features at which the highest accuracy is obtained for the proposed system is 79.6% for SVM and 75% for k-NN, with LPCC features optimized at Rank-17 and Rank-12, respectively.

Furthermore, the performance of the proposed SER system is compared with different works in Table 4.4 for EMO-DB and Table 4.5 for IEMOCAP in terms of the number of optimized features and classification accuracy performance measures.

Table 4.4: Comparison of existing SER works with the proposed SER system for EMO-DB

Approaches		**No. of Features**	**Classification Accuracy**
Chen et al. (2016) [90]		72	77.74%
Zhang et al. (2013) [99]		9	80.85%
Zhang et al. (2013) [91]		11	73.9%
Yan et al. (2013) [98]		13	79.23%
Kuchibhotla et al. (2016) [96]		12	88.1%
Özseven (2019) [94]		304	84.07%
Sun et al. (2019) [95]		500	86.86%
Proposed SER	**SVM**	**18**	**73.65%**
		56	**90.12%**
	k-NN	**19**	**88.2%**
		60	**89.23%**

In [90], semi-NMF with k-means clustering initialization was used to transform feature sets, which were further combined with the original dataset to obtain a total of 72 features for SER obtaining 77.74% accuracy. In [91], [96], [98], [99], different optimizing and feature selection techniques, namely enhanced kernel isometric mapping, the modified supervised locally linear embedding algorithm, sparse partial least squares regression, sequential floating forward selection, the scaled conjugate gradient, and principal component analysis, were used for improving the classification accuracy by reducing the feature set dimension. However, the classification accuracy obtained with the proposed SER system is higher than the other methods with 90% (approx.) using both classification techniques for the EMO-DB database. In [94], [95], a new statistical feature selection and Fisher feature selection were used to select the most optimal feature sets, but still, these techniques have

lower performance both in terms of complexity, i.e. the number of features, and classification accuracy compared to the proposed SER system.

Table 4.5: Comparison of existing SER works with the proposed SER system for IEMOCAP

Approaches		**No. of Features**	**Classification Accuracy**
Sahu et al. (2017) [88]		100	58.38%
Latif et al. (2018) [89]		128	56.42%
Proposed SER	**SVM**	**17**	**79.6%**
		53	**83.2%**
	k-NN	**12**	**75%**
		60	**78%**

Likewise, in [88], [89], variational and adversarial auto-encoders were used for feature optimization and the performance was lower than that of the proposed SER system for the IEMOCAP database with a classification accuracy of 79.6% for 17 features and 83.2% for 53 features using the SVM classifier and 75% for 12 features and 78% for 60 features using the K-NN classifier.

4.4 Summary

In the proposed SER system, the semi-NMF feature optimization technique with SVD initialization is employed to optimize the MFCC, LPCC, and TEO-AutoCorr features. The proposed SER system performance is evaluated in the presence of the EMO-DB and IEMOCAP databases. The 5-fold cross-validation scheme is used for the classification of the emotions using k-nearest neighborhood and support vector machine classifiers. The cross-validation phenomenon considers the entire dataset for both training and testing to avoid overfitting problems. The optimal rank is chosen for semi-NMF depending on the database and classification technique used for MFCC, LPCC, and TEO-AutoCorr features. The

combination of these optimized feature sets is used in the proposed SER system to achieve the highest classification accuracies of 90.12% and 89.3% for the EMO-DB database and 83.2% and 78% for the IEMOCAP database, with SVM and k-NN classification techniques respectively. It is evident from the results that the proposed SER system outperforms the baseline, i.e. the SER system without optimization, and also the existing literature works. Here, the semi-NMF technique employed for feature optimization is a transformation technique and due to this there is a lack of data interpretability and also, even after optimization, the time taken by the classifier to train and validate the data is very high. To retain the data interpretability of feature data even after optimization, feature selection algorithms can be used.

Chapter 5

Speech Emotion Recognition using Unsupervised Feature Selection

In this chapter, an SER system is proposed with unsupervised feature selection algorithms to optimize 1582 INTERSPEECH 2010 Paralinguistic and 20 Gammatone Cepstral Coefficients (GTCC) feature set. The Support Vector Machine (SVM) classifier using Linear and Radial Basis Function (RBF) kernels with 10-Fold Cross-Validation and Hold-Out Validation is used to classify the emotions. The proposed SER system is validated with performance metrics - Computational Time and Classification accuracy. The significant contributions of this chapter are:

i) Using the UFSOL and FSASL unsupervised feature selection algorithms for feature selection which have not yet been explored for SER.

ii) Proposing a Subset Feature Selection (SuFS) algorithm to further improve the performance of the proposed SER system by selecting the subset of features after UFSOL and FSASL feature selection.

5.1 Motivation:

The SER system developed using the semi-NMF feature optimization technique lacks data interpretability. This is because the Semi-NMF algorithm transforms the feature data into another domain. But in machine learning data interpretability is very important. Interpretability is a circumstance where humans can predict the result of a model reliably. The higher the data is interpretable, the better is the model. If the decisions of a model are easily comprehendible by humans then the model is better interpretable. The feature selection algorithms can also be used for feature optimization. In feature selection, the prominent features that contribute to emotion recognition are selected without transforming the data. Thus, by using feature selection for feature optimization in the SER system, there is no lack of interpretability.

5.2 Proposed Speech Emotion Recognition System using Unsupervised Feature Selection Algorithms

The proposed SER system is developed using unsupervised feature selection algorithms and with higher frame size. When a higher frame size is used for speech pre-processing in speech emotion recognition, the variation in classification accuracy is not much [116]. In the proposed SER system, after the feature extraction, the unsupervised feature selection algorithms, i.e., UFSOL and FSASL are used individually to select the most prominent from the original feature set as shown in figure 5.1. By using a higher frame size, the feature data dimension is reduced which further reduces the computation time in the SER system. The speech signal is initially passed through a pre-emphasis filter to boost the energy in their higher frequencies which are attenuated during the speech signal production from the vocal tract [23].

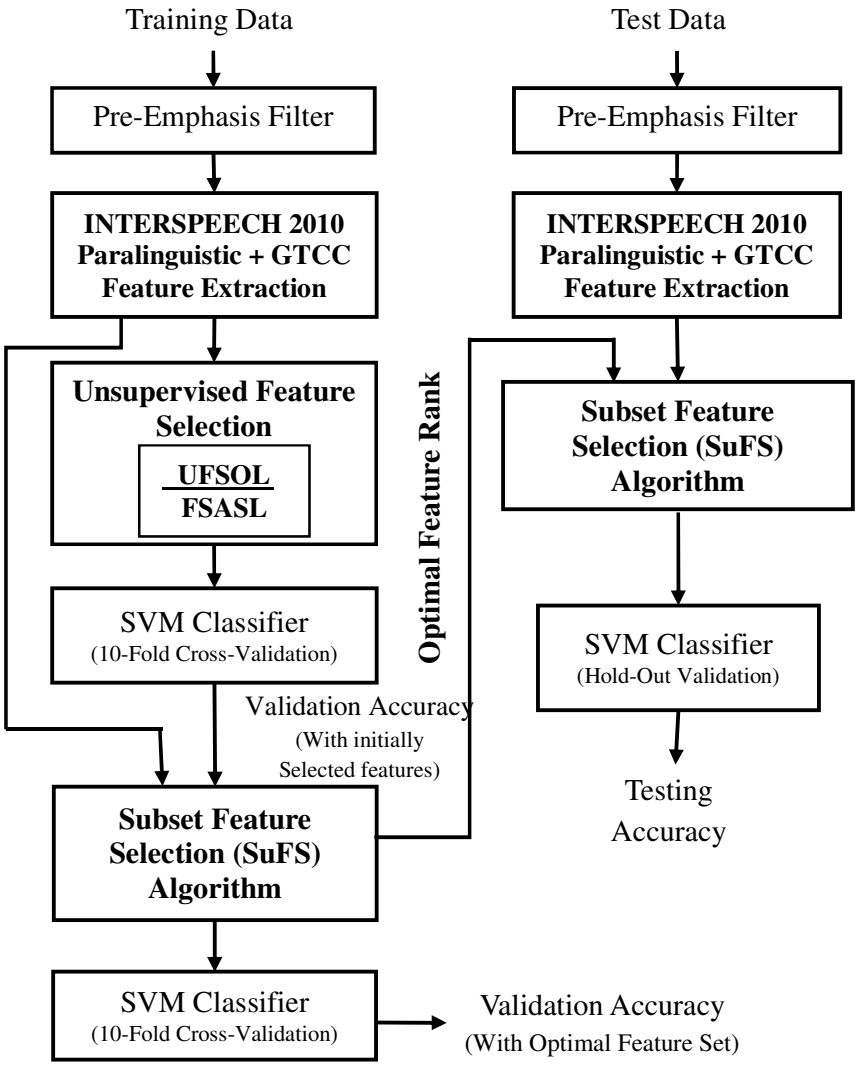


Figure 5.1: Proposed SER system using unsupervised feature selection

5.2.1 Feature Extraction

In this work, the combination of INTERSPEECH 2010 paralinguistic features and Gammatone Cepstral Coefficients (GTCC) are used as features.

5.2.1.1 INTERSPEECH 2010 Paralinguistic Feature Set

The INTERSPEECH 2010 paralinguistic challenge set consists of 1582 features with a four-set of features combined [77]. The set of 1582 features are extracted using openSMILE toolkit from a single speech signal [124]. 'IS10paraling:conf" is the configuration file to obtain the feature set. These features, along with the description are shown in table 5.1.

Table 5.1: INTERSPEECH 2010 paralinguistic feature set

Descriptors	**Functionals**
PCM Loudness	Position – max./ min.
MFCC [0-14]	Arithmetic mean, Standard Deviation
Log Mel Freq. Band [0-7]	Skewness, Kurtosis
LSP Frequency [0-7]	Linear regression coefficient
F0 by Sub-Harmonic Sum.	Linear regression error
F0 Envelope	Quartile
Voicing Probability	Quartile range
Jitter Local	Percentile
Jitter DDP	Percentile range
Shimmer Local	Up-level time

5.2.1.2 Gammatone Cepstral Coefficients (GTCC):

The Gammatone filter takes its name from the impulse response, which is the product of a Gamma distribution function and a sinusoidal tone centred at the frequency, being computed as [125]:

$$g(t) = Kt^{(n-1)}e^{-2\pi Bt}\cos(2\pi f_c t + \phi) \tag{5.1}$$

where '$g(t)$' is the impulse response of gammatone filter; 'K' is the amplitude factor; 'n' is the filter order; 'f_c' is the central frequency in Hz; 'ϕ' is the phase shift; 'B' is the duration of the impulse response ($B = 1.019 \times ERB(f_c)$).

ERB is the equivalent rectangular bandwidth i.e., $ERB(f) = 24.7 + 0.108f$. The centre frequency f_c of each gammatone filter is equally spaced on ERB scale, i.e.,

$$f_c = ERBS^{-1}(ERBS(f_{low}) + \frac{ERBS(f_{high} - f_{low})}{N}) \quad (5.2)$$

where, $ERBS(f) = 21.4 \log_{10}(1 + 0.00437f)$.

The fourth-order gammatone filter is similar to a human auditory model, therefore $n = 4$. Here, $f_{low} = 62.5\ Hz$, $f_{high} = 3400\ Hz$ and N is the number of gammatone filters i.e., 20. After obtaining the gammatone filter coefficients the cepstral analysis is applied to these, obtaining a total of '20' gammatone cepstral coefficients using the gammatone filter.

5.2.2 Unsupervised Feature Selection Algorithms

The unsupervised feature selection algorithms, i.e., UFSOL and FSASL, which are not yet explored for SER so far, are used in this work. Apart from this, a novel Subset Feature Selection algorithm is modelled by the results obtained after using UFSOL and FSASL algorithms to improve the performance of the SER system further. The entire set of 1602 features is given to the feature selection algorithms to select the most prominent features, as shown in figure 5.1.

5.2.2.1 Unsupervised Feature Selection with Ordinal Locality (UFSOL):

Consider $X = [x_1, .., x_d] \in \mathbb{R}^{m \times d}$ as the initial feature matrix with *d* speech signals and '*m*' number of features. Generally the regularized regression, feature selection is formulated as [126]:

$$min_W \|W^T X - H\|_F^2 + \delta \|W\|_{2,q} \tag{5.3}$$

where, $W \in \mathbb{R}^{m \times d_2}$ $(m > d_2)$ is a projection matrix/ feature selection matrix; $l_{2,q}$-norm (q is typically set to 0 or 1) assures the sparseness in rows of '*W*'; $H = [h_1, .., h_d] \in \mathbb{R}^{d_2 \times d}$is a target matrix in this unsupervised feature selection algorithm.

Whereas, '*H*' is a label matrix in the case of supervised multi-class data. In this work, the bi-orthogonal semi Non-negative Matrix Factorization (NMF) is used to decompose *H* into two new matrices *i.e.*, $H \cong UV$ with $V \geq 0, VV^T = I$ and $U^T U = I$.

If the feature set selected for the original sample x_i is supposed to be $y_i = W^T x_i$, then $Y = W^T X$. According to the principle of "*ordinal locality preserving*", given a triplet (x_i, x_u, x_v) comprised of x_i and its neighbors x_u and x_v, their corresponding feature groups also form a triplet (y_i, y_u, y_v). Let the distance metric be denoted by $dist(.,.)$. The feature selection holds *ordinal locality preserving* if the following condition is preserved:

i.e., if $dist(x_i, x_u) \leq dist(x_i, x_v)$, then $dist(y_i, y_u) \leq dist(y_i, y_v)$.

Based on this, the appropriate feature group for each data point is identical to optimizing the following ordinal locality preserving loss function over a collection of triplets as below:

$$\max_Y \sum_{i=1}^{d} \sum_{u \in N_i} \sum_{v \in N_i} A_{uv}^{i} [dist(y_i, y_u) - dist(y_i, y_v)] \tag{5.4}$$

where, N_i is a set of sequence numbers indicating the 'k' nearest neighbors of x_i; A^i denotes an antisymmetric matrix with $(u, v)^{th}$ element, the $dist(x_i, x_u) - dist(x_i, x_v)$. If the weighting matrix is denoted as $C \in \mathbb{R}^{d \times d}$ then

$$C_{i,j} = \begin{cases} \sum_{u \in N_i} A_{uj}^{i} & , j \in N_i \\ 0 & , j \notin N_i \end{cases} \tag{5.5}$$

From (5.5), the equation (5.4) is equivalent to

$$\min_Y \sum_{i=1}^{d} \sum_{j=1}^{d} C_{ij} dist(y_i, y_j) \tag{5.6}$$

The squared Euclidean distance is used to establish each pairwise distance. The loss function of ordinal locality preserving can be written accordingly as $\min_Y \sum_{i=1}^{d} \sum_{j=1}^{d} C_{ij} \|y_i - y_j\|_2^2$, which has an equivalent compact matrix form: $\min_Y Tr(YLY^T)$ as well as $\min_W Tr(W^T XLX^T W)$ by substituting $Y = W^T X$. From these considerations, (5.3) can be formulated as,

$$\min_{W,U,V} F = \{\|W^T X - UV\|_F^2 + \delta \|W\|_{2,1} + \rho Tr(W^T XLX^T W)\},$$

$$s.t. W^T W = I, V \geq 0, VV^T = I \tag{5.7}$$

where δ and ρ are scalar constants that control the relativeness of corresponding terms.

According to half-quadratic theory, for a fixed 't', there is a conjugate function $\psi(.)$, with $\sqrt{t^2+\varepsilon} = inf_{r\epsilon R}\left\{\frac{r}{2}t^2 + \psi(r)\right\}$. The infimum could be reached at $r = {}^{1}/_{\sqrt{t^2+\varepsilon}}$. With this, (5.7) can be optimized by minimizing its augmented function $\hat{F}$ as below:

$$\min_{W,U,V,R} \hat{F} = \left\{\|W^T X - H\|_F^2 + \delta \sum_{i=1}^{m}\left\{\frac{R_{ii}}{2}\|W_i\|_2^2 + \psi_i(R_{ii})\right\} + \rho Tr(W^T X L X^T W)\right\}$$
$$\text{, s.t. } W^T W = I\ V \geq 0, VV^T = I \tag{5.8}$$

where, 'R' is a $m \times d_2$ diagonal matrix storing the auxiliary variables and $\{\psi_i\}_{i=1}^{m}$ are conjugate functions. i.e.,

$$\min_{W,U,V} F(W,U,V) = \min_{W,U,V,R} \hat{F}(W,U,V,R) \tag{5.9}$$

The minimization of $\hat{F}(W,U,V,R)$ is as shown below:

i) The diagonal elements of 'R' are updated in parallel:

$$R_{ii} = {}^{1}/_{\sqrt{\|W_i\|_2^2 + \varepsilon}} \tag{5.10}$$

ii) To solve (5.8), (U,V) is updated for fixed W by applying orthogonal Semi-NMF on projected data i.e., feature selection matrix Y =W^TX. The orthogonal semi-NMF problem $\min_{U,V}\|Y - UV\|_F^2,\ s.t. V \geq 0, VV^T$ is equivalent to relaxed k-means clustering. The zero gradient condition $U = W^T X V^T$is attained by updating (U,V) using k-means clustering.

The algorithm to solve (5.8) is as below:

Algorithm 5.1: The algorithm to solve (5.8)

Input: Data matrix $X = [x_1, .., x_d] \in \mathbb{R}^{m \times d}$;

Number of each sample's nearest neighbors k;

Parameters d_2, c, δ and ρ.

Solution:

1: Compute C via (5.7) and its corresponding Laplacian matrix L;

2: Initialize $\mathrm{W}^{(0)}$ with d_2 different columns randomly selected $\mathrm{d}_1 \times \mathrm{d}_1$ identity matrix, $\mathrm{t} = 0$;

3: while not convergence do

4: $\mathrm{t} \leftarrow \mathrm{t} + 1$;

5: Update $\mathrm{R}^{(\mathrm{t})}$ via (5.10);

6: Update $\mathrm{U}^{(\mathrm{t})}$ and $\mathrm{V}^{(\mathrm{t})}$ by K-means;

7: Update $\mathrm{W}^{(\mathrm{t})}$ by Eigen decomposition;

8: end while

Output: $W \rightarrow$ Feature Selection matrix; $V \rightarrow$ cluster indicator matrix.

All the above steps are updated until convergence as summarized in Algorithm 1. W is the resultant feature selection matrix.

iii) W is updated with (U, V) fixed, substitute $U = W^T X V^T$ in f(W, U, V) and the objective function $\min_{W^T W=I} Tr(W^T G W)$ is solved by applying Eigen decomposition on $G = \frac{\delta}{2} R + X(\rho L + I - V^T V)X^T$. The optimal W comprises d_2 Eigen vectors corresponding to the smallest Eigen values of d_2.

5.2.2.2 Feature Selection with Adaptive Structure Learning (FSASL):

In this algorithm, consider the feature set as $X \in \mathbb{R}^{d \times m}$, where '*d*' corresponds to the speech files dimension and '*m*' as the total number of features. The regularization parameters considered are α, β, γ, μ and these are used for error reconstruction and balancing sparsity of global and local structure learning. Further, '*c*' is considered as the dimension for the optimized data, resulting in a feature optimization set that belongs to $\mathbb{R}^{d \times c}$. The FSASL is achieved using the general equation [127]:

$$\min_{Z,S,P}(\|Z^T X - Z^T X S\|^2 + \alpha\|S\|_1) + \beta \sum_{q,r}^{m} \left(\left\| Z^T x_q - Z^T x_r \right\|^2 P_{qr} + \mu {P_{qr}}^2 \right) + \gamma \|Z\|_{21}$$
$$\text{subject to } S_{qr} = 0, \bar{P} 1_m = 1_m, \bar{P} \geq 0, Z^T X X^T Z = \bar{I}; \quad (5.11)$$

where, $X \rightarrow$ Feature set of input; $x \rightarrow$ specific row of data matrix;

Algorithm 2: FSASL Algorithm

Input:

Input feature set as $X \in \mathbb{R}^{d \times m}$;

'*d*' → speech files dimension;

'*m*' → total number of features.

Solution:

For each data sample x_q, consider x_q that has probability *P(q,r)* for all the data points $\{x_r\}_{r=1}^{m}$ of

$S \rightarrow$ Weight matrix corresponding to data matrix;

$s \rightarrow$ specific row corresponding to weight matrix;

$Z \rightarrow$ resultant feature selection matrix after transformation;

The problem of optimization as in (5.11) results in the dissimilar variables $(S, P \; and \; Z(t))$ into a set of single variable sub-problems and solved as follows:

i) With *P* and *Z* as constants, the matrix *S* is solved. The q^{th} column of *S* is updated for each value of 'q' by solving the problem:

$$\min_{s_q}(\|x_q{}' - X^q s_q\|^2 + \alpha|s_q|), \text{ s.t. } S_{qq} = 0 \tag{5.12}$$

X' and x' are the transpose matrices of *X* and *x*.

ii) Solving for *P* by keeping the *S* and *Z* constant. For each *q*, update q^{th} column of *P* by solving the problem

$$\min_{W,S,P} \sum_{q,r}^{m} \left(\|x_q{}' - x_r{}'\|^2 P_{qr} + \mu {P_{qr}}^2\right), \text{ s.t. } 1_m^T p_q = 1, P_{qr} \geq 0 \tag{5.13}$$

Denote $A \in \mathbb{R}^{m \times m}$ be a square matrix with $A_{qr} = -\frac{1}{2\mu}\left\|x_q{}' - x_r{}'\right\|^2$, then the above problem can be written as:

$$\min_{p_q^T} \frac{1}{2}\left\|p_q^T - a_q^T\right\|^2, \text{ s.t. } p_q^T 1_m = 1, 0 \leq p_{qr}^T \leq 1 \tag{5.14}$$

where $p'(t)$ is the q^{th} row of P.

iii) The graph laplacian i..e., *L=Ls + β (Lp)* is computed; where,

$$L_P = D_P - (P + P^T)/2 \tag{5.15}$$

where, D_P is a diagonal matrix whose ith diagonal element is $\sum_r \frac{P_{qr}+P_{rq}}{2}$

$$L_S = (I - S)(I - S)^T \tag{5.16}$$

iv) With P and S as constants, solve the matrix Z i.e., a feature selection matrix, using the following equation:

$$\min_Z \ Tr(Z^T X L X^T Z^T) + \gamma\|Z\|_{21}, \text{ s.t. } Z^T X X^T Z = I \tag{5.17}$$

The (5.17) is modified keeping the i^{th} diagonal element equal to $\frac{1}{2\|z_q^t\|^2}$ as:

$$\min_W Tr(Z^T X(L + \gamma D_{Z^t})X^T Z, \text{ s.t. } Z^T X X^T Z = I \tag{5.18}$$

where D_{Z^t} is the diagonal matrix and Z^t is the t^{th} estimation.

'Z' is the resultant optimal solution obtained from the Eigen vectors of the 'c' Eigen values that are the smallest, derived from the eigen-problem:

$$X(L + \gamma D_{Z^t})X^T Z = \Lambda X X^T Z \tag{5.19}$$

where Λ is a diagonal matrix whose diagonal elements are Eigen values.

Output: Based on $||z_q||_2$ ($q = 1, ..., d$) , all the d features are sorted in descending order such that prominent 'k' ranked features are selected to obtain the Z matrix.

The resultant is the Z as the feature selection matrix. Both the FSASL and UFSOL algorithms, rearrange the original feature set accordingly, as per their prominence with the ranks of the corresponding algorithms. Later, the rearranged feature sets are given to the SVM classifier to perform emotion classification or prediction.

5.2.2.3 Subset Feature Selection (SuFS):

After the unsupervised feature selection, a novel Subset Feature Selection algorithm is introduced upon the UFSOL and FSASL algorithms. To further reduce the dimension of the feature set without affecting the accuracy of the SER system, i.e., to obtain a better accuracy with a reduced feature set. The SuFS algorithm is discussed in Algorithm 5.3. The data of the results obtained from the SER system UFSOL and FSASL algorithm with the accuracies for the corresponding features are considered in the algorithm. The inputs to this algorithm are the original feature set, the rank of the UFSOL/ FSASL algorithm and the highest accuracies obtained for different number of features using the UFSOL/FSASL algorithm.

Algorithm 5.3: Subset Feature Selection (SuFS)

Input: Ranking vector *r* based on Unsupervised Feature Selection;

Original Feature Vector '*F*' (1602 features);

Accuracy Vector (*a*) with accuracies based on the ranking of various features using Feature Selection algorithm;

l = number of features at which the highest accuracy is obtained using UFSOL or FSASL.

Solution:

1: Initialize sub-rank (sr) with a(1) (since, first accuracy value is always > 0)

2: Initialize h=2

for g=0:1:l

if a(g+1) > a(g)

sr(h) = r(g+1)

update h ← h+1

end

3: for i=0:1:len(sr)

sf(g) = F(:, sr(g))

end

Output: *Subset of original feature vector (sf)*

In the first step, the rank of the UFSOL or FSASL algorithm i.e., the features at which the highest accuracy is obtained using these algorithms are considered as the initial sub-rank (sr). In the second step, the 'sr' is updated such that the accuracy is being increased continuously. If there is any decrease in the accuracy value from the previous accuracy for a particular feature, then the corresponding feature is excluded from the initially selected features. This iteration continues, until all the features that give less accuracy are removed and the final 'sr' is the updated SuFS rank with the most optimal features. In the final step, the subset of the original feature vector (sf) is obtained from the initial UFSOL/FSASL selected features.

The proposed SuFS depends on the ranking vector (i.e., prominence of the features) and the validation accuracy obtained from the features selected from UFSOL and FSASL algorithms. The ranking vector is according to d_2 smallest Eigen values of UFSOL algorithm and d smallest Eigen values of FSASL algorithm.

The SuFS algorithm is applied to the features selected by UFSOL and FSASL to obtain the *'sf'* feature vector. Further, the subset of features, i.e., features obtained from UFSOL-SuFS and FSASL-SuFS are given to the SVM classifier for both validation and testing.

5.3 Simulation Results and Performance Evaluation

In the proposed SER system, the 1602 INTERSPEECH Paralinguistic and GTCC features are extracted from the speech signal. This huge set of features is fed to the UFSOL and FSASL algorithms for feature selection. In this work, the support vector machine (SVM) classification technique with Linear and Radial Basis Function (RBF) kernels using Hold-Out and 10-fold Cross-Validation are used for emotion classification. Initially, the speech signal database is divided into training and testing datasets. 80% of the dataset is considered for training and 20% for testing for hold-out validation. In this work, the 10-fold cross-validation

schema is used to train and test the accuracy of the proposed SER system. Hence, the entire dataset is randomly split into 10 parts, among that 9 parts are used for training the classifier (SVM), and testing is carried out on the hold-out or test data, i.e., the tenth part. This process is repeated in 10 folds, i.e., 10 times, until the entire dataset is completely trained. The experimental analysis is carried out using EMO-DB and IEMOCAP databases.

The performance of the proposed SER system is evaluated using the machine learning performance metric, i.e., the Classification Accuracy. In this work, the 10-fold cross-validation and Hold-Out Validation are used to train and test the accuracy of the proposed SER system. All the simulations are carried out in a Computer with Intel(R) Xeon(R) CPU E3-1220 v3 of 3.10 GHz 64-bit processor with 16 GB RAM. The simulation parameters used in the proposed SER system are shown in table 5.2.

Table 5.2: Simulation Parameters of Proposed SER System using unsupervised FS

Parameters	**Specifications**
Pre-Emphasis Filter	Coefficient, a = 0.97
Frame Size/ Length	4096 samples
Frame Overlap	1024 samples
Type of Window	Hamming
Gammatone Filter	Filter order = 4 f_{low} = 62.5Hz, f_{high} = 3400Hz Number of gammatone filters = 20
Validation	Hold-Out Validation (80/20) 10-Fold Cross-Validation
Support Vector Machine	Kernel = Linear, Radial Basis Function

To select the first prominent features which give the highest accuracy, to select the initial feature set, the feature selection matrix of both UFSOL and FSASL algorithms are given to the SVM classifier as shown in Figure 5.1.

Figures 5.2 and 5.3 show the variation of classification accuracy with the number of features using FSASL and UFSOL feature selection for EMO-DB and IEMOCAP. For EMO-DB, using FSASL the highest validation accuracy of 86% is obtained for 600 features and 85% validation accuracy for 500 features with UFSOL. For IEMOCAP, for 1250 features the highest accuracy of 71.4% using FSASL and 72% using UFSOL is obtained.

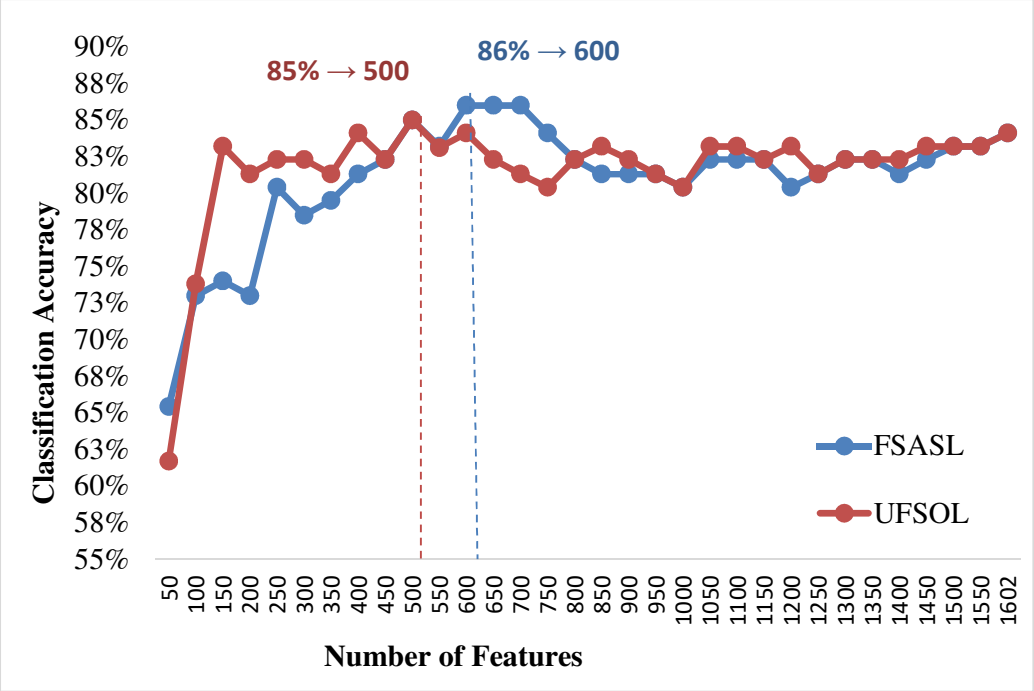


Figure 5.2: Variation of classification accuracy in proposed SER system with FSASL and UFSOL feature selection using SVM classifier (10-fold cross-validation) with EMO-DB

The best GTCC features selected for EMO-DB are GTCC [1] using FSASL and FSASL-SuFS, GTCC [2] using UFSOL and UFSOL-SuFS. While, for IEMOCAP, GTCC [1-

20] i.e., the entire GTCC feature set is selected using FSASL, GTCC [1, 2, 4-7, 11] using FSASL-SuFS, GTCC [1-19] using UFSOL and GTCC [3-5, 7, 9, 10, 12-19] using UFSOL-SuFS. The best INTERSPEECH Paralinguistic 2010 features selected by each of the feature selection algorithms that are considered in the proposed SER are shown in Table I of Appendix I.

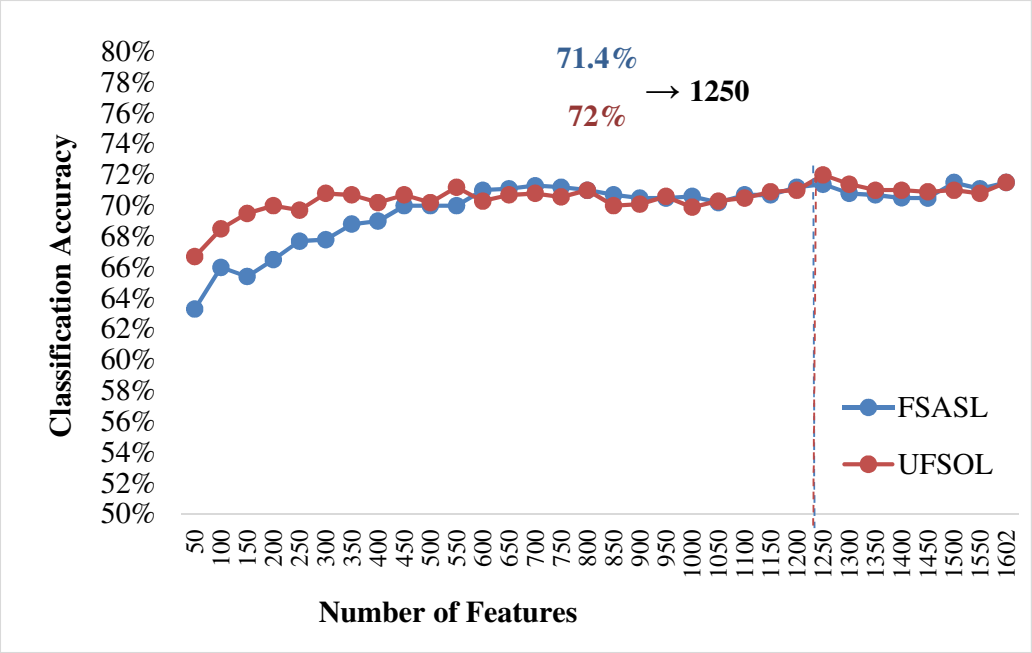


Figure 5.3: Variation of classification accuracy in proposed SER system with FSASL and UFSOL feature selection using SVM classifier (10-fold cross-validation) with IEMOCAP

It is evident from Figures 5.2 and 5.3, even with initially selected features using UFSOL and FSASL algorithms, the SER accuracy is not increasing. Therefore, still, the feature selection is possible from initially chosen features. Hence, the SuFS algorithm is applied after UFSOL and FSASL feature selection to acquire better accuracy with less number of features. The initially selected features are fed to the SuFS algorithm to reduce further the number of features acquiring the best performance. Later, the highest prominent

features selected by SuFS are fed to the SVM classifier with Linear and RBF kernels for emotion classification.

The performance of the proposed SER system with different feature selection algorithms is compared with the baseline SER system (without feature selection) using SVM classifier with Linear and RBF kernels using hold-out validation and 10-fold cross-validation are as shown in tables 5.3 and 5.4 in terms of classification accuracy and validation (or) testing time. The processing time is calculated per training/ testing fold.

Table 5.3: Performance comparison of baseline and proposed SER systems for EMO-DB and IEMOCAP databases with SVM classifier using hold-out validation

Database	**Method**	**No. of Features**	**Linear Kernel**			**RBF Kernel**		
			Training	**Testing**		**Training**	**Testing**	
			Time (sec)	**Time (sec)**	**Acc (%)**	**Time (sec)**	**Time (sec)**	**Acc (%)**
EMO-DB	Baseline	1602	6.4	0.17	84.1	1.3	0.22	76.6
	UFSOL	500	0.22	0.05	85	0.41	0.06	75.7
	FSASL	600	0.28	0.06	86.8	0.47	0.07	75.7
	UFSOL-SuFS	450	0.21	0.043	84.9	0.57	0.08	74.8
	FSASL-SuFS	350	0.165	0.032	86	0.29	0.04	77.6
IEMOCAP	Baseline	1602	39.35	5.4	56.05	46.4	10.9	71
	UFSOL	1250	35	5.4	56.05	29.4	5.9	70.9
	FSASL	1250	34.1	5.3	57.3	34.6	7.7	70
	UFSOL-SuFS	800	21.2	3.4	60.6	14	2.9	77.5
	FSASL-SuFS	650	24.6	2.9	59.7	21	4.1	70.4

From the results shown in tables 5.3 and 5.4, it is clear that the SVM with Linear kernel gives better classification for EMO-DB data and with RBF kernel in the case of IEMOCAP data. Table 5.4 shows the hold-out validation results for EMO-DB and IEMOCAP database. For EMO-DB, the highest testing accuracy of 86% with the lowest computational time for training and testing, i.e., 0.165 and 0.032 seconds using the FSASL-SuFS algorithm. Similarly, for IEMOCAP database, the highest testing accuracy and lowest computational time of 14 and 2.9 seconds for training and testing is 77.5% using the UFSOL-SuFS algorithm.

Table 5.4: Performance comparison of the baseline and proposed SER system for EMO-DB and IEMOCAP databases using SVM classifier with 10-fold cross-validation

Database	**Method**	**No. of Features**	**Linear Kernel**		**RBF Kernel**	
			Time (sec)	**Acc(%)**	**Time (sec)**	**Acc(%)**
EMO-DB	Baseline	1602	3.2	85(±0.8)	12.13	81(±1.5)
	UFSOL	500	2.5	86(±1.0)	4.07	78((±1.5)
	FSASL	600	1.86	85(±1.3)	5.1	78((±1.4)
	UFSOL-SuFS	450	1.71	84(±0.8)	5.4	81((±1.4)
	FSASL-SuFS	350	1.4	85(±0.8)	2.68	78((±1.3)
IEMOCAP	Baseline	1602	304.9	58(±0.3)	430	69(±0.4)
	UFSOL	1250	289.4	58(±0.5)	310	69(±0.4)
	FSASL	1250	277.5	59(±0.5)	309	69(±0.4)
	UFSOL-SuFS	800	216.7	57(±0.5)	125.5	77(±0.4)
	FSASL-SuFS	650	199	58(±0.4)	199.8	69(±0.4)

In table 5.4, for EMO-DB database, using SVM with Linear kernel the 10-fold cross-validation accuracy of the baseline SER system without feature selection is 85(±0.8)% with 1602 features. After applying the feature selection algorithms, the dimension of the feature set is reduced. The proposed SER system achieves an accuracy of 86(±)% using UFSOL with selected 500 features and 85(±1.3)% using FSASL with 600 selected features. The SuFS

algorithm is applied on these selected features of UFSOL and FSASL, thus reducing the number of features and acquiring the accuracy of 85(±1.5)% for UFSOL-SuFS with 450 features and 85(±0.8)% for FSASL-SuFS with 350 features.

Similarly, for IEMOCAP database from the results shown in table 5.4 using SVM with RBF kernel, the 10-fold cross-validation accuracy of the baseline SER system without feature selection is 69(±0.4)% with 1602 features. After feature selection, the proposed SER system achieves an accuracy of 69(±0.4)% using UFSOL and FSASL with selected 1250 selected features. The accuracy with UFSOL-SuFS is 77(±0.4)% with 800 features and 69(±0.4)% for FSASL-SuFS with 650 features. The confusion matrices with individual accuracy of each emotion of EMO-DB and IEMOCAP database using the proposed SER system with baseline, FSASL, UFSOL, FSASL-SuFS and UFSOL-SuFS are shown in tables 5.5 to 5.14.

Table 5.5: Confusion matrix of baseline SER system for EMO-DB

Emotion	**Ang**	**Anx**	**Bor**	**Dis**	**Hap**	**Neu**	**Sad**
Ang	**94.1%**	0	0	0	5.9%	0	0
Anx	5.5%	**77.7%**	0	0	16.8%	0	0
Bor	0	0	**79%**	0	0	10.5%	10.5%
Dis	0	0	0	**75%**	0	12.5%	12.5%
Hap	6.2%	0	0	6.3%	**87.5%**	0	0
Neu	0	5.9%	5.9%	0	0	**88.2%**	0
Sad	0	0	16.7%	0	0	0	**83.3%**

Table 5.6: Confusion matrix of proposed FSASL based SER system for EMO-DB

Emotion	**Ang**	**Anx**	**Bor**	**Dis**	**Hap**	**Neu**	**Sad**
Ang	**92.4%**	0	0	0	7.6%	0	0
Anx	5.5%	**84.5%**	0	0	10%	0	0
Bor	0	0	**79%**	0	0	10.5%	10.5%
Dis	0	0	0	**87.5%**	0	12.5%	0
Hap	12.5%	0	0	0	**87.5%**	0	0
Neu	0	0	5.9%	0	0	**94.1%**	0
Sad	0	0	16.7%	0	0	0	**83.3%**

Table 5.7: Confusion matrix of proposed UFSOL based SER system for EMO-DB

Emotion	**Ang**	**Anx**	**Bor**	**Dis**	**Hap**	**Neu**	**Sad**
Ang	**94.1%**	0	0	0	5.9%	0	0
Anx	5.6%	**83.3%**	0	0	11.1%	0	0
Bor	0	0	**84.2%**	0	0	5.3%	10.5%
Dis	0	0	0	**75%**	0	25%	0
Hap	25%	0	0	0	**75%**	0	0
Neu	0	0	5.9%	0	0	**94.1%**	0
Sad	0	0	16.7%	0	0	0	**83.3%**

Table 5.8: Confusion matrix of proposed FSASL-SuFS based SER system for EMO-DB

Emotion	**Ang**	**Anx**	**Bor**	**Dis**	**Hap**	**Neu**	**Sad**
Ang	**90.3%**	0	0	0	9.7%	0	0
Anx	9%	**82%**	0	9%	0	0	0
Bor	0	0	**84.2%**	0	0	5.3%	10.5%
Dis	12.5%	0	0	**87.5%**	0	0	0
Hap	12.5%	0	0	0	**87.5%**	0	0
Neu	0	0	5.9%	0	0	**94.1%**	0
Sad	0	0	16.7%	0	0	0	**83.3%**

Table 5.9: Confusion matrix of proposed SER system with UFSOL-SuFS for EMO-DB

Emotion	**Ang**	**Anx**	**Bor**	**Dis**	**Hap**	**Neu**	**Sad**
Ang	**96%**	0	0	0	4%	0	0
Anx	5.6%	**83.3%**	0	0	1.1%	0	0
Bor	0	0	**76.4%**	0	0	11.8%	11.8%
Dis	0	12.5%	0	**75%**	0	0	12.5%
Hap	6.2%	0	0	6.3%	**87.5%**	0	0
Neu	0	4.1%	5.9%	0	0	**90%**	0
Sad	0	0	16.7%	0	0	0	**83.3%**

Table 5.10: Confusion matrix of baseline SER system for IEMOCAP

Emotion	**Ang**	**Hap**	**Neu**	**Sad**
Ang	**84%**	1.3%	13.8%	0.9%
Hap	10.8%	**18%**	50.4%	20.8%
Neu	4.5%	3.5%	**80.2%**	11.8%
Sad	2.2%	1.3%	24.1%	**72.4%**

Table 5.11: Confusion matrix of proposed FSASL based SER system for IEMOCAP

Emotion	**Ang**	**Hap**	**Neu**	**Sad**
Ang	**79.1%**	1.8%	17.8%	1.3%
Hap	10%	**19%**	45.9%	25.1%
Neu	4.5%	2%	**83%**	10.5%
Sad	3.5%	1.3%	24.2%	**71%**

Table 5.12: Confusion matrix of proposed UFSOL based SER system for IEMOCAP

Emotion	**Ang**	**Hap**	**Neu**	**Sad**
Ang	**80.5%**	1.3%	16.9%	1.3%
Hap	11.7%	**18.1%**	46.8%	23.4%
Neu	3.1%	2.5%	**83.3%**	11.1%
Sad	3.5%	1.7%	23.2%	**71.6%**

Table 5.13: Confusion matrix of proposed FSASL-SuFS based SER system for IEMOCAP

Emotion	**Ang**	**Hap**	**Neu**	**Sad**
Ang	**85.5%**	5.8%	6.9%	1.8%
Hap	6.3%	**20.7%**	55%	18%
Neu	4.2%	3.5%	**80.9%**	11.4%
Sad	4%	2.5%	24.6%	**68.9%**

Table 5.14: Confusion matrix of proposed UFSOL-SuFS based SER system for IEMOCAP

Emotion	**Ang**	**Hap**	**Neu**	**Sad**
Ang	**97.3%**	0.9%	0.9%	0.9%
Hap	8.1%	**22.6%**	50.4%	18.9%
Neu	1.7%	2.4%	**86.5%**	9.4%
Sad	1.3%	2.6%	22.8%	**73.3%**

From the results, it is clearly understood that by using the unsupervised feature selection and inducing SuFS algorithm upon UFSOL and FSASL techniques, the proposed SER system provides improved accuracy with less computational complexity.

Further, the performance of the proposed SER system is compared with the different works in table 5.14 for EMO-DB and IEMOCAP databases in terms of the Classification Accuracy performance metric. It is evident that the proposed SER system upon using the feature selection process provided improved performance compared to the rest of the SER systems in the literature.

Table 5.15: Comparison of proposed unsupervised FS based SER system with existing works

Methods		**EMO-DB**	**IEMOCAP**
Chen et al. 2016 [90]		77.4%	-
Zhang et al. 2013 [91]		80.85%	-
Zhang and Zhao 2013 [99]		78.5%	-
Yan et al. 2013 [98]		79.23%	-
Gudmalwar et al. 2019 [100]		75.32%	-
Ozseven 2019 [94]		84.07%	-
Sun et al. 2019 [95]		86.86%	-
Huang et al. 2015 [101]		71.16%	-
Sahu et al. 2018 [88]		-	58.38%
Latif et al. 2017 [89]		-	56.42%
Jiang et al. 2019 [128]		-	64%
Proposed SER System	**FSASL**	**86(±1.0)%**	**69(±0.4)%**
	UFSOL	**85(±1.3)%**	**69(±0.4)%**
	FSASL-SuFS	**85(±1.5)%**	**77(±0.4)%**
	UFSOL-SuFS	**85(±0.8)%**	**69(±0.4)%**

5.4 Summary

In this chapter, the unsupervised feature selection techniques UFSOL and FSASL are employed to optimize the combination of INTERSPEECH 2010 Paralinguistic and GTCC features. Also, a novel SuFS algorithm is proposed upon the UFSOL and FSASL techniques to reduce further the feature dimension acquiring the comparable performance in the proposed SER system. The performance of the proposed SER system is analyzed with EMO-DB and IEMOCAP databases using SVM classifier with Linear and RBF kernels. 10-fold Cross-validation scheme is used to train the feature sets so as to consider the entire dataset for both training and testing to avoid the over-fitting problem and Hold-Out validation scheme to test

the performance of the proposed SER system with new data. The proposed SER system for EMO-DB data achieves highest classification accuracy using SVM with Linear kernel with 86% using FSASL and 85% using UFSOL, FSASL-SuFS and UFSOL-SuFS methods. Similarly, the highest classification accuracy for IEMOCAP database is obtained using SVM classifier with RBF kernel with 77% using FSASL-SuFS and 69% using the rest of the methods respectively. It is clearly evident from the results that the proposed SER system outperforms the baseline, i.e., the SER system without feature selection and also with the existing literature works.

Chapter 6

Noise Robust Speech Emotion Recognition using PNCC Features and NMF De-Noising

In this chapter, a noise robust SER system is proposed to improve the SER classification accuracy in noisy environments. The power normalized cepstral coefficient (PNCC) features provide better SER accuracy in noisy conditions. Therefore, an SER system is proposed with the combination of INTERSPEECH 2010 Paralinguistic Feature Set, Gammatone Cepstral Coefficients (GTCC) and PNCC speech features. The SER system proposed uses unsupervised feature selection algorithms to select the best features from the huge feature set and the Support Vector Machine (SVM) classifier using Linear and Radial Basis Function (RBF) kernels for emotion classification. The proposed SER system is evaluated both in clean and noisy speech scenarios. The proposed SER system shows almost the same performance as a clean speech environment for the noisy speech signal with SNR values greater than 15dB because of the use of GTCC and PNCC features. For the noisy speech with lower SNR values, the dense Non-Negative Matrix Factorization (denseNMF) is used to de-noise the noisy speech signal. Later, the feature extraction and feature selection are performed acquiring noise robustness.

The significant contributions of the proposed work in this chapter are:

i) Using the PNCC features along with the INTERSPEECH 2010 Paralinguistic Feature Set and GTCC for acquiring improved SER accuracy.

ii) Applying the dense Non-negative Matrix Factorization (denseNMF) for de-noising the noisy speech signal corrupted with different noise signals at various Signal-to-Noise Ratio (SNR) levels to acquire noise robustness in SER.

6.1 Motivation

The SER systems so far have been developed in a clean speech environment without noise interference. In a real-time scenario, different types of noises can corrupt the speech signal and it becomes vulnerable, losing the speech intelligibility. Due to this, there can be a reduction in SER accuracy. Hence, there is a need to make the SER system robust to noisy conditions and improve SER accuracy. Therefore to overcome this disadvantage, in the proposed SER system, a noise robust PNCC feature set and NMF de-noising technique in the speech pre-processing stage are used to achieve noise robustness.

6.2 Proposed Speech Emotion Recognition System using Power Normalized Cepstral Coefficients and DenseNMF De-noising

In this work, the proposed SER system is evaluated initially using clean speech data and later the speech data corrupted with different noisy signals at various SNR levels. Initially, the speech signal is pre-processed using a pre-emphasis filter and then the speech features are extracted. After feature extraction, the unsupervised feature selection algorithms, i.e., UFSOL and FSASL are used individually to select the most prominent features from the original feature set as shown in figure 6.1.

6.2.1 Database

In the proposed work, EMO-DB and IEMOCAP datasets are considered for the SER analysis. For noise analysis, the Aurora noisy database is used in this work [110]. In this database, the noises have been recorded at different places like Suburban train, Crowd of people (babble), Car, Exhibition hall, Restaurant, Street, Airport, Train station. The Additive White Gaussian Noise (AWGN) and noises of the Aurora database (airport, babble, car, station and street) are used for the analysis in the proposed work.

6.2.2 Feature Extraction

In SER, the emotional relevant speech features are extracted from speech signals using the feature extraction process [8]. In order to obtain the emotional contents from a speech signal, a particular set of features can be extracted by applying various signal processing techniques. In this work, the feature fusion of INTERSPEECH 2010 paralinguistic features, Gammatone Cepstral Coefficients (GTCC) and Power Normalized Cepstral Coefficients (PNCC) is used.

6.2.2.1 Power Normalized Cepstral Coefficients (PNCC):

The Power Normalized Cepstral Coefficients (PNCC) are robust against different sources of environmental disturbances such as background additive noise, linear channel distortion, and reverberation [129]. 13 PNCC features are extracted which are useful in noise SER analysis [106].

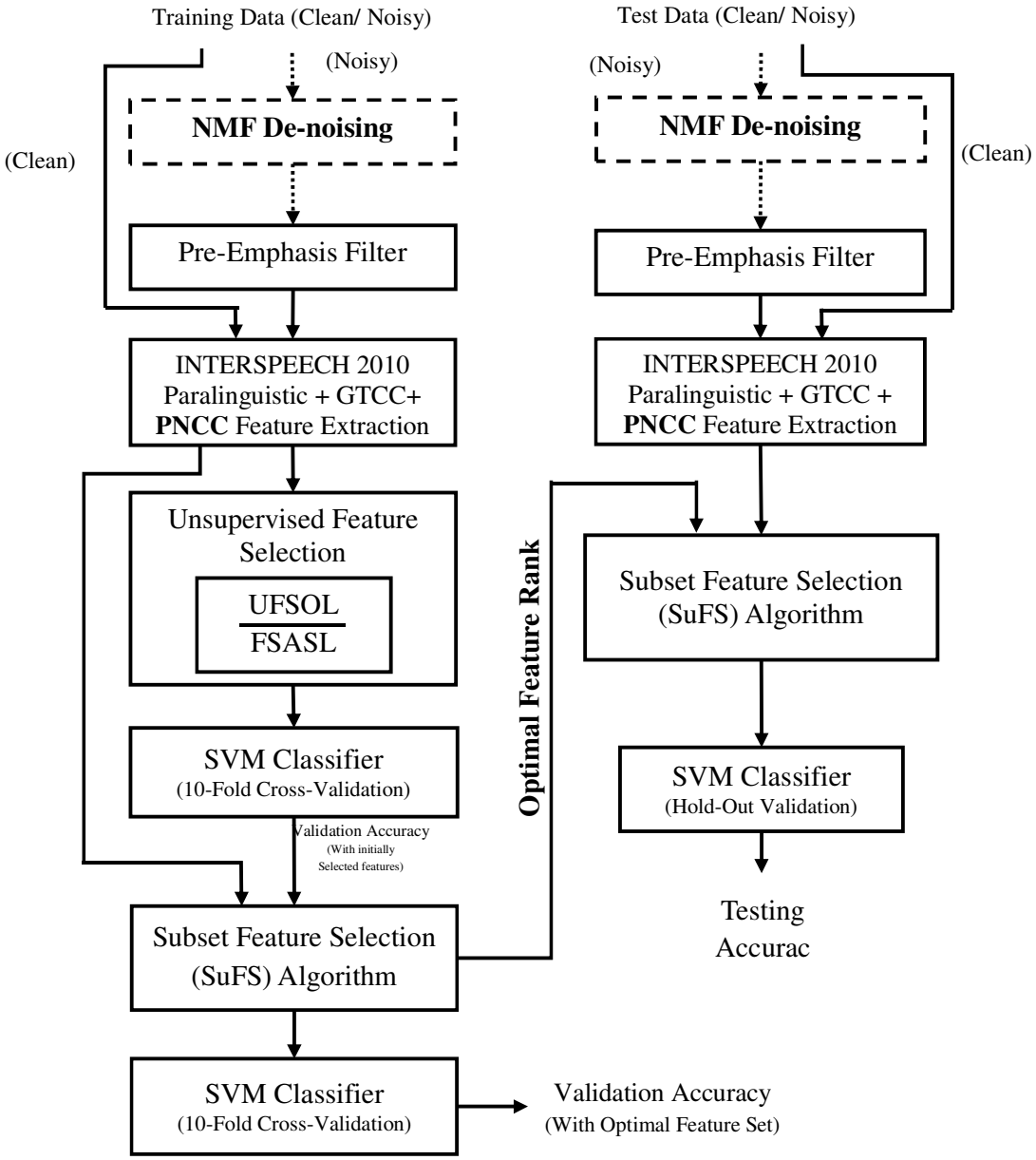


Figure 6.1: Proposed Noise Robust Speech Emotion Recognition system

6.2.3 Unsupervised Feature Selection

The unsupervised feature selection algorithms, i.e., UFSOL and FSASL, which are not yet explored for SER so far, are used in this work. Apart from this, a novel Subset Feature Selection algorithm is modelled by the results obtained after using UFSOL and FSASL algorithms to improve the performance of the SER system further. These feature selection algorithms select the most prominent features among the huge set of 1615 features as shown in figure 6.1. The UFSOL and FSASL algorithms are discussed in chapter 5.

6.2.4 Noise Analysis

In a real-time environment, the speech signal is vulnerable to different noisy conditions. Due to this, the performance of the SER system is degraded. To overcome this disadvantage, speech de-noising before emotion recognition is an efficient way. Many speech de-noising techniques have been developed so far for enhancing the speech quality without any disturbances [130], [131]. The speech de-noising techniques are designed to remove the noise added through any source from corrupted speech. Spectral subtraction is one of the simple and most widely used methods for estimating noise spectral profile in the magnitude domain which is subtracted from the speech segments. But the spectral subtraction method is restricted to quasi-stationary signal analysis. Recently, Non-negative Matrix Factorization (NMF) has become the popular technique that makes use of speech signal representations, specifically in the perspective of blind source separation [132]. In this method, the magnitude spectrogram 'H' is factorized to two non-negative matrices 'H = UV'. Here the magnitude spectral profiles are the columns of matrix 'U' and the gain coefficients are present in 'V'.

6.2.4.1 Noise Analysis

In this work, a denseNMF based speech de-noising is used [133]. Consider U_{sp} as the speech and U_{no} as the noise matrix dictionaries. From speech production mechanism, the

speech signal is the convolutive model of the excitation source $e(t)$ and vocal tract filter a(t) in time-domain [134]:

$$sp(t) = a(t) * e(t) \tag{6.1}$$

The excitation signal $e(t)$ is represented as [135]:

$$e(t) = \sum_{k=1}^{p} c_k \exp(\Im k\bar{\omega}(t)) \tag{6.2}$$

where, $\bar{\omega}(t)$ is the fundamental frequency.

From [135], the input speech spectrogram $H_{sp} \in \mathbb{R}_{+}^{K\times T}$ holds:

$$\begin{cases} H_{sp} = U_{sp}V_{sp} \\ d_f^{sp} = E_f C a_f, \qquad f = 1,2,\dots m_{sp} \end{cases} \tag{6.3}$$

Each column d_j^{sp} of matrix $U_{sp} \in \mathbb{R}_{+}^{K\times m_{sp}}$represents one harmonic column, in which isolated harmonics are placed in columns of matrices $E_f \in \mathbb{R}_{+}^{K\times p}$, weighted by constant amplitude matrix $C = diag(c_1,\dots c_p)$. The representation coefficients $a_f \in \mathbb{R}_{+}^{p}$ and gain matrix V_{sp} are needed to be defined.

Similarly, noise spectrogram $M_{no} \in \mathbb{R}_{+}^{K\times T}$ holds

$$\begin{cases} H_{no} = U_{no}V_{no} \\ d_f^{no} = Ob_f, \qquad f = 1,2,\dots m_{no} \end{cases} \tag{6.4}$$

With $U_{no} \in \mathbb{R}_{+}^{K \times m_{no}}$ represents noise dictionary with atoms d_f^{no}, $0 \in \mathbb{R}_{+}^{K \times z}$ contains noise spectral shapes combined with unknown coefficients $b_f \in \mathbb{R}_{+}^{z}$ to produce a noise model.

Constrained speech and noise dictionaries U_{sp} and U_{no} are combined into matrix $U = [U_{sp} U_{no}]$, and gain matrix $V = \left[V_{sp}^T V_{no}^T\right]^T$ is randomly initialized. The denseNMF optimization task is adopted as in [133]:

$$\begin{cases} U_{KD}(H \parallel UV + \lambda \parallel V \parallel_1 + \sigma \sum_f \parallel a_f \parallel_2^2 \rightarrow min \\ d_f^{sp} = \Psi_f a_f, \parallel a_f \parallel_1 \quad f = 1,2, \dots m_f \end{cases} \tag{6.5}$$

The following rules are also derived from multiplicative updates for l_1-normalized coefficients $\tilde{a}_f = a_f, \parallel a_f \parallel_1$:

$$a_f \leftarrow \tilde{a}_f . \left(\mathbb{1}_f \tilde{a}_f^T \Psi_f^T \mathbb{1} \bar{r}_f^T + \Psi_j^T \frac{\boldsymbol{H}}{\boldsymbol{UV}} \bar{r}_j^T + \sigma \mathbb{1}_j \tilde{a}_r^T \tilde{a}_r \right) / \left(\Psi_r^T \mathbb{1} \bar{x}_r^T + \mathbb{1}_r \tilde{a}_r^T \Psi_r^T \frac{\boldsymbol{H}}{\boldsymbol{UV}} \bar{x}_r^T + \sigma \tilde{a}_r \right) \tag{6.6}$$

$$V \leftarrow V . \left((U^T \frac{\boldsymbol{H}}{\boldsymbol{UV}}) / (Q^T \mathbb{1} + \lambda) \right) \tag{6.7}$$

where, $\mathbb{1}_f$ indicates the vector of all-ones of the same size as a_f.

Using iterative updates (6.7) all representation coefficients and gain matrices are obtained. Thus, obtaining the resultant de-noised signal.

6.3 Simulation Results and Performance Evaluation

The proposed SER system is validated in both clean and noisy environments. Initially, the feature selection algorithms are adopted on the clean speech database for feature dimension reduction and later, the different clean speech data is corrupted using different types of noises and checked for noise robustness. From the clean speech signal data, a combination of 1582 INTERSPEECH 2010 paralinguistic, 20 GTCC and 13 PNCC features are extracted. This huge set of features is fed to the UFSOL and FSASL algorithms for feature selection. In this work, the support vector machine (SVM) classification technique with Linear and Radial Basis Function (RBF) kernels, using hold-out and 10-fold cross-validation, is used for emotion classification.

For hold-out validation, the speech signal database is divided into training and testing datasets, 80% of the dataset is considered for training and 20% for testing. Cross-validation is a resampling method used to analyze the machine learning algorithms '*k*' (here, *k*=10) number of folds on a small dataset. With the entire set of data, equally sized '*k*' folds or groups are formed randomly. In this set of groups, one fold is used for validation and the rest are used to fit the model on the '*k*-1' folds. Here, '*k*' is considered to be 10 i.e., 10-fold cross-validation is adopted. Accordingly, the entire data is divided randomly into 10 folds, where the first fold is hold-out for validation and the rest 9 folds are to train the SVM classifier. This process is repeated in 10 folds, i.e., 10 times, until the entire dataset is completely trained. In this work, the 10-fold cross-validation and hold-out validation are used in the analysis of the proposed SER system.

All the simulations are carried out in a Computer with Intel(R) Xeon(R) CPU E3-1220 v3 of 3.10 GHz 64-bit processor with 16 GB RAM.

6.3.1 Proposed SER analysis in Clean Environment

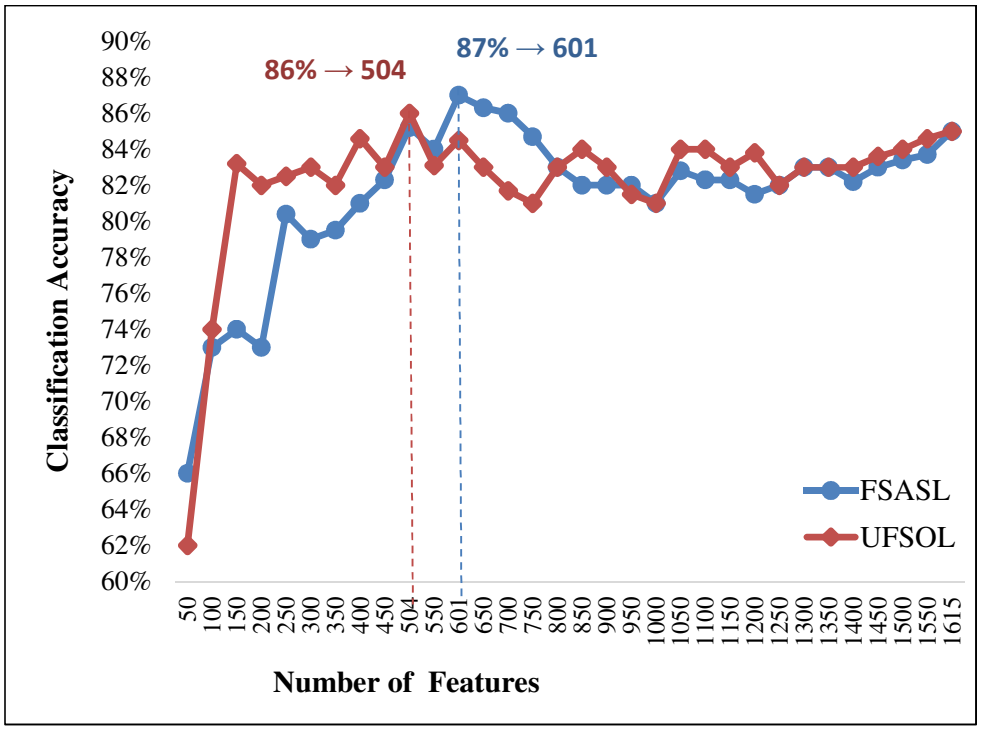


Figure 6.2: Performance Variation of Proposed SER system with FSASL and UFSOL feature selection using SVM classifier (10-fold cross-validation) with EMO-DB database

To select the first prominent features which give the highest accuracy, to select the initial feature set, the feature selection matrix of both UFSOL and FSASL algorithms are given to the SVM classifier as shown in figure 6.1. Figures 6.2 and 6.3 show the variation of classification accuracy with the number of features using FSASL and UFSOL feature selection for EMO-DB and IEMOCAP. For EMO-DB, using FSASL the highest validation accuracy of 87% is obtained for 601 features and 86% validation accuracy for 504 features with UFSOL. For IEMOCAP, for 1254 features the highest accuracy of 72% by using FSASL and 73% by using UFSOL is obtained.

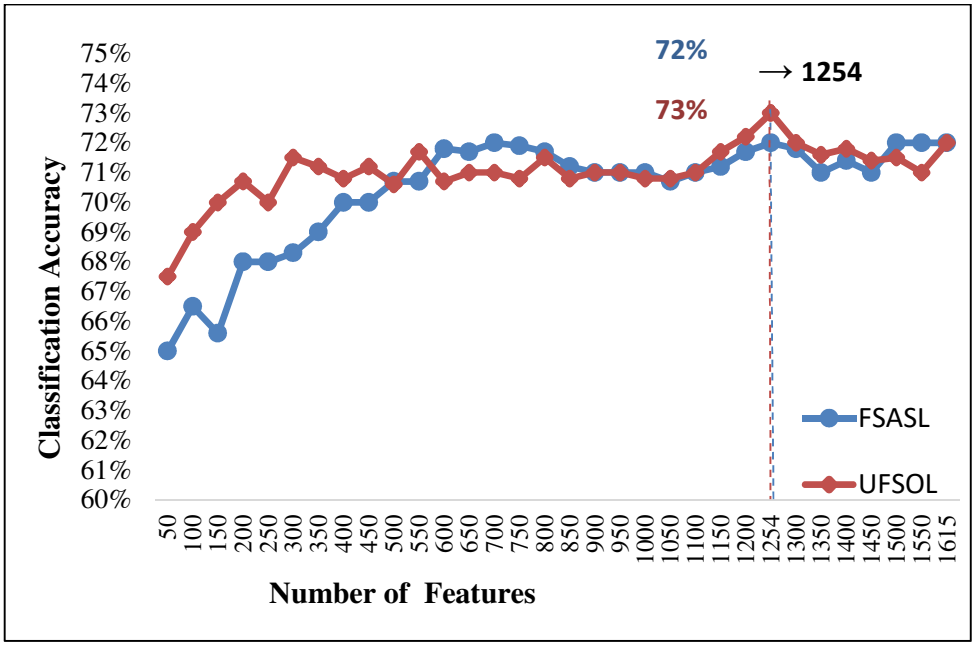


Figure 6.3: Performance Variation of Proposed SER system with FSASL and UFSOL feature selection using SVM classifier (10-fold cross-validation) with IEMOCAP database

It is evident from figures 6.2 and 6.3, even with initially selected features using UFSOL and FSASL algorithms, the SER accuracy is not increasing. Therefore, still, the feature selection is possible from initially chosen features. Hence, the SuFS algorithm is applied after UFSOL and FSASL feature selection to acquire better accuracy with less number of features.

The initially selected features of UFSOL and FSASL algorithms are fed to the SuFS algorithm to reduce further the number of features for acquiring the best performance. These features are given to the SVM classifier with Linear and RBF kernels for emotion classification. The classification accuracy and computational time i.e, training and testing time are used for the evaluation of the proposed SER system performance analysis.

Table 6.1: Performance Comparison of Baseline and Proposed SER system for EMO-DB & IEMOCAP Database using SVM classifier with 10-fold Cross-Validation

Database	Method	No. of Features	Linear Kernel		RBF Kernel	
			Time (sec)	Acc (%)	Time (sec)	Acc (%)
EMO-DB	Baseline	1615	3.5	85(±1.0)	12.3	82(±0.3)
	UFSOL	504	2.6	86(±0.8)	4	79((±0.7)
	FSASL	601	2	**87(±1.3)**	5.15	79((±0.5)
	UFSOL-SuFS	**454**	1.7	**85(±0.5)**	5.5	82((±0.4)
	FSASL-SuFS	**352**	1.4	**85(±1.0)**	2.9	79((±0.2)
IEMOCAP	Baseline	1615	306	59(±1.2)	432	72(±1.3)
	UFSOL	1254	289	59(±0.7)	311	72(±1.0)
	FSASL	1254	278.2	60(±1.3)	308	73(±0.8)
	UFSOL-SuFS	**802**	217	58(±0.4)	126	**77(±1.4)**
	FSASL-SuFS	**653**	199.4	58(±1.2)	199.9	**71(±1.2)**

Tables 6.1 and 6.2 show the performance comparison of the proposed SER system with feature selection algorithms and baseline SER system without feature selection using SVM classifier with Linear and RBF kernels using 10-fold cross-validation and hold-out validation in terms of classification accuracy and validation (or) testing time. Tables 6.1 and 6.2 show the simulation results of the proposed SER system using the SVM classifier for the EMO-DB and IEMOCAP database. From the results, it can be clearly understood that for the EMO-DB database, better performance is achieved upon using the SVM classifier with Linear Kernel, and RBF kernel for the IEMOCAP database.

Table 6.2: Performance Comparison of Baseline and Proposed SER system for EMO-DB & IEMOCAP Database using SVM classifier with Hold-Out Validation

Database	Method	No. of Features	Linear Kernel			RBF Kernel		
			Training	Testing		Training	Testing	
			Time (sec)	Time (sec)	Acc (%)	Time (sec)	Time (sec)	Acc (%)
EMO-DB	Baseline	1615	6.82	0.17	84.5	1.5	0.3	77.2
	UFSOL	504	**0.29**	**0.05**	**85.5**	**0.4**	**0.1**	**76.3**
	FSASL	601	**0.31**	**0.06**	**87**	**0.5**	**0.09**	**76.7**
	UFSOL -SuFS	**454**	0.22	0.043	85.3	0.6	0.07	75.4
	FSASL-SuFS	**352**	0.17	0.032	86.3	0.3	0.05	78
IEMOCAP	Baseline	1615	40	5.5	57	48.2	11	71.4
	UFSOL	1254	35.2	5.45	56.82	**30**	5.9	71
	FSASL	1254	34	5.3	58	**36**	7.8	70.5
	UFSOL-SuFS	**802**	22.3	3.5	61	14.2	3	77.8
	FSASL-SuFS	**653**	25	3	60.7	21	4	71.2

From the results shown in tables 6.1 and 6.2, it is clear that the SVM with Linear Kernel gives better classification for EMO-DB data and with RBF kernel in the case of IEMOCAP data. For the EMO-DB database, using SVM with Linear kernel the 10-fold cross-validation accuracy of baseline SER system without feature selection is 85($\pm$1.0)% with 1615 features. After applying the feature selection algorithms, the dimension of the feature set is reduced. The proposed SER system achieves an accuracy of 86($\pm$0.8)% using UFSOL with selected 504 features and 87($\pm$1.3)% using FSASL with 601 selected features. The SuFS algorithm is applied on these selected features of UFSOL and FSASL, thus reducing the number of features and acquiring the accuracy of 85($\pm$0.5)% for UFSOL-SuFS with 454 features and 85($\pm$1.0)% for FSASL-SuFS with 352 features.

Similarly, for the IEMOCAP database using SVM with RBF kernel, the 10-fold cross-validation accuracy of the baseline SER system without feature selection is 72(±1.3)% with 1615 features. After feature selection, the proposed SER system achieves an accuracy of 72(±1.0)% and 73(±0.8)% using UFSOL and FSASL with selected 1254 selected features. The accuracy with UFSOL-SuFS is 77(±1.4)% with 802 features and 71(±1.2)% for FSASL-SuFS with 653 features.

Table 6.3: Performance comparison of proposed work with the existing works

Methods	**EMO-DB**	**IEMOCAP**
Chen et al. 2016 [90]	77.4%	-
Zhang et al. 2013 [91]	80.85%	-
Zhang & Zhao 2013 [99]	78.5%	-
Yan et al. 2013[98]	79.23%	-
Gudmalwar et al. 2019 [100]	75.32%	-
Ozseven 2019 [94]	84.07%	-
Sun et al. 2019 [95]	86.86%	-
Huang et al. 2015 [101]	71.16%	-
Sahu et al. 2017 [88]	-	58.38%
Latif et al. 2017 [89]	-	56.42%
Jiang et al. 2019 [128]	-	64%
SER using FSASL **(Proposed)**	**87(±1.3)%**	**73(±0.8)%**
SER using UFSOL **(Proposed)**	**86(±0.8)%**	**72(±1.0)%**
SER using FSASL-SuFS **(Proposed)**	**85(±1.0)%**	**71(±1.2)%**
SER using UFSOL-SuFS **(Proposed)**	**85(±0.5)%**	**77 (±1.4)%**

Table 6.2 shows the hold-out validation results for EMO-DB and IEMOCAP database. For EMO-DB, the highest SER testing accuracy of 87% using FSASL and a comparable accuracy of 86.3% with the lowest training time of 0.17 seconds is achieved using the

FSASL-SuFS algorithm. Similarly, for the IEMOCAP database, the highest testing accuracy achieved with the lowest training time of 14.2 seconds is 77.8% using the UFSOL-SuFS algorithm.

From the results, it is clearly understood that by using the unsupervised feature selection and inducing the SuFS algorithm upon UFSOL and FSASL techniques, there is an improvement in the accuracy of the proposed SER system with less computational complexity.

Further, the performance of the proposed SER system is compared with the different works in Table 6.3 for EMO-DB and IEMOCAP databases in terms of the Classification Accuracy performance metric. The performance of the proposed SER system upon using the feature selection process provided improved performance compared to the rest of the SER systems in the literature.

6.3.2 Proposed SER Analysis in Noisy Environment

The clean speech of the EMO-DB and IEMOCAP database are corrupted with the different noises from the Aurora database [110]. The original clean speech is corrupted with different kinds of noises i.e., airport, babble, car, station, street and white at SNR levels from -5dB to 20dB.

Figures 6.4 to 6.7 show the performance of the proposed SER system with unsupervised feature selection in presence of noisy speech using hold-out and 10-fold cross-validation accuracies. The initially selected feature labels of the proposed SER system using UFSOL, FSASL, FSASL-SuFS and UFSOL-SuFS are used for the noisy analysis.

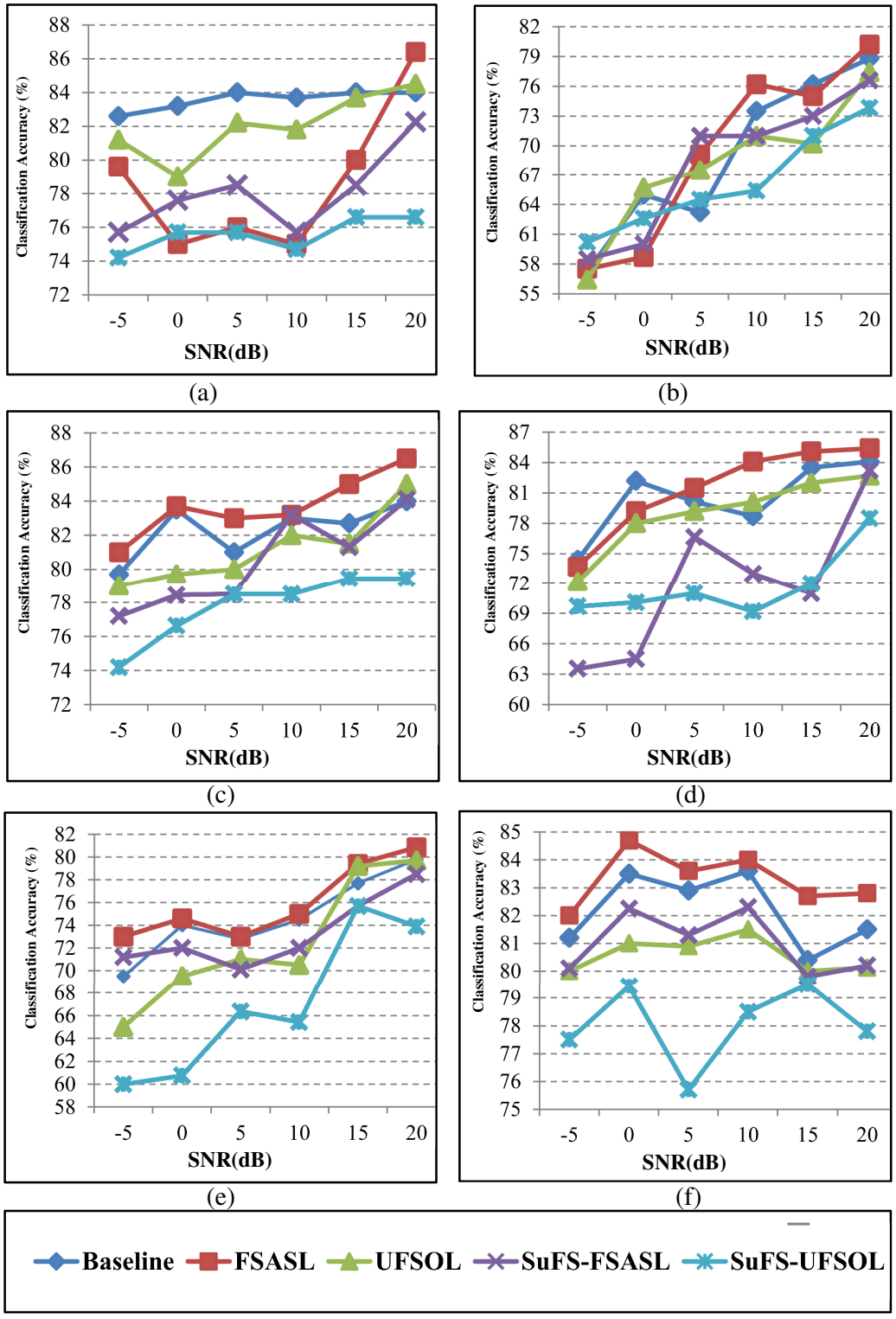


Figure 6.4: Hold-out validation accuracy variations of the proposed SER system for EMO-DB noisy data with different noises (a) airport (b) babble (c) car (d) station (e) street (f) white

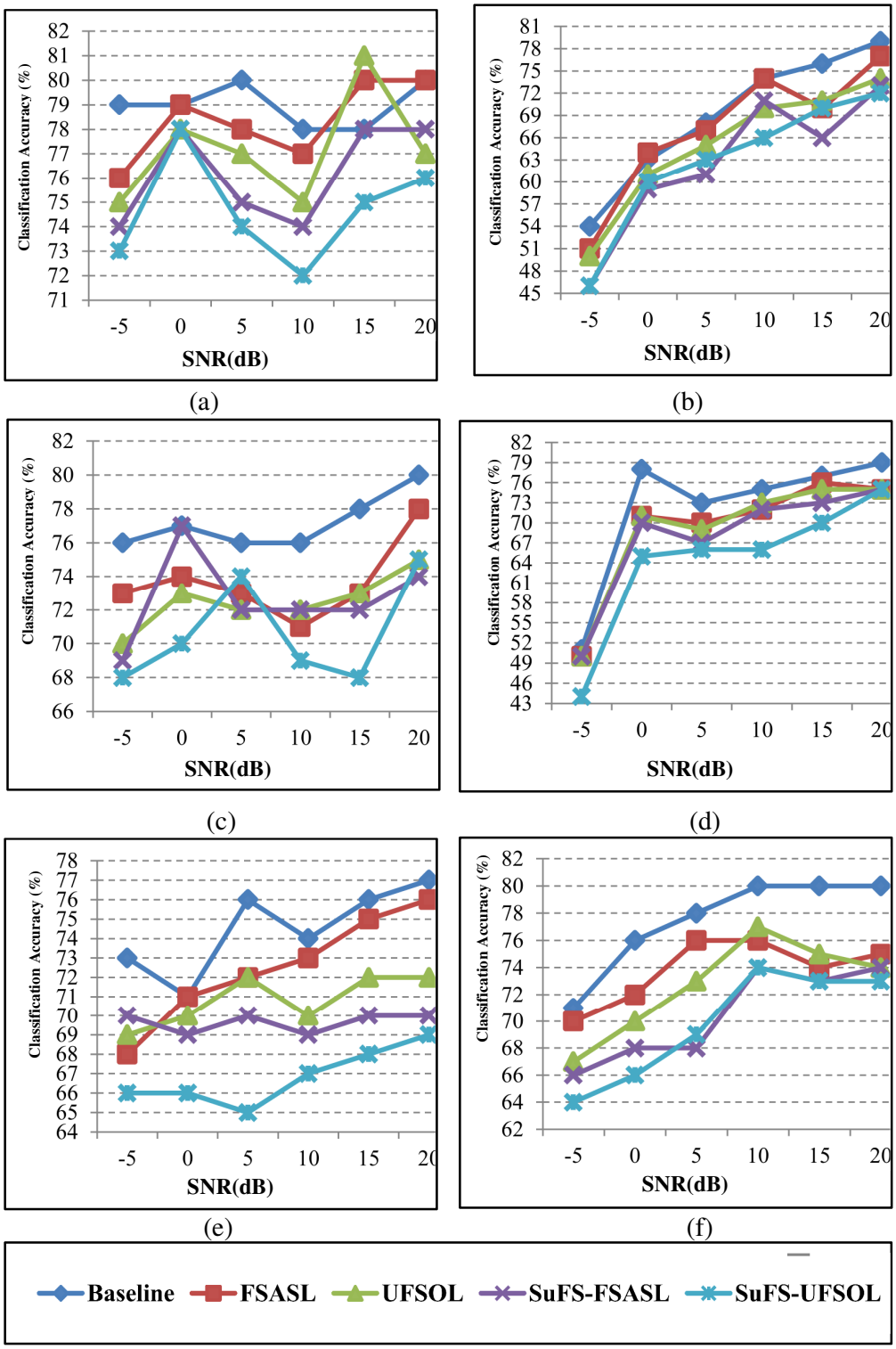


Figure 6.5 10-Fold Cross-Validation accuracy variations of the proposed SER system for EMO-DB noisy data with different noises (a) airport (b) babble (c) car (d) station (e) street (f) white

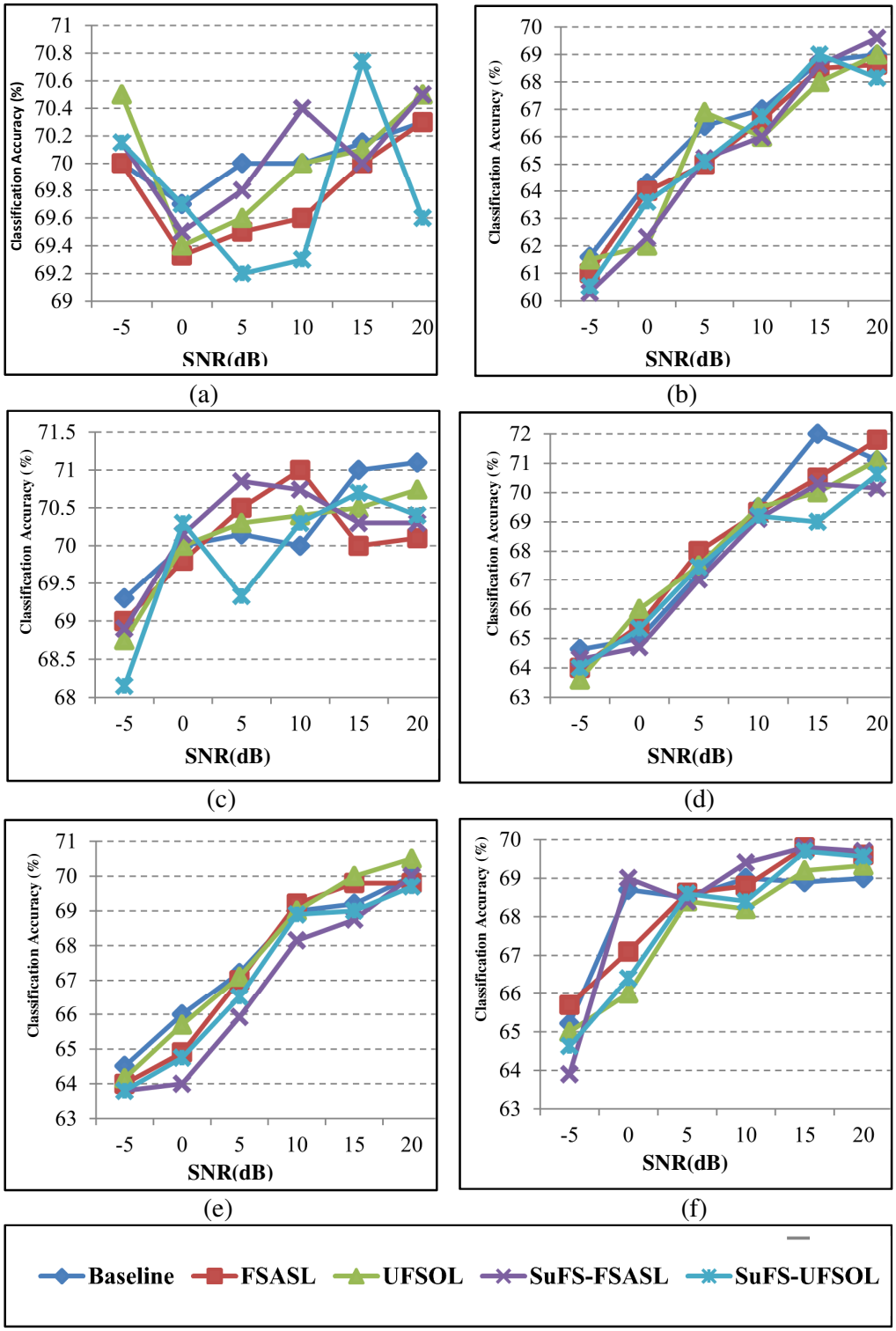


Figure 6.6: Hold-out validation accuracy variations of the proposed SER system for IEMOCAP noisy data with different noises (a) airport (b) babble (c) car (d) station (e) street (f) white

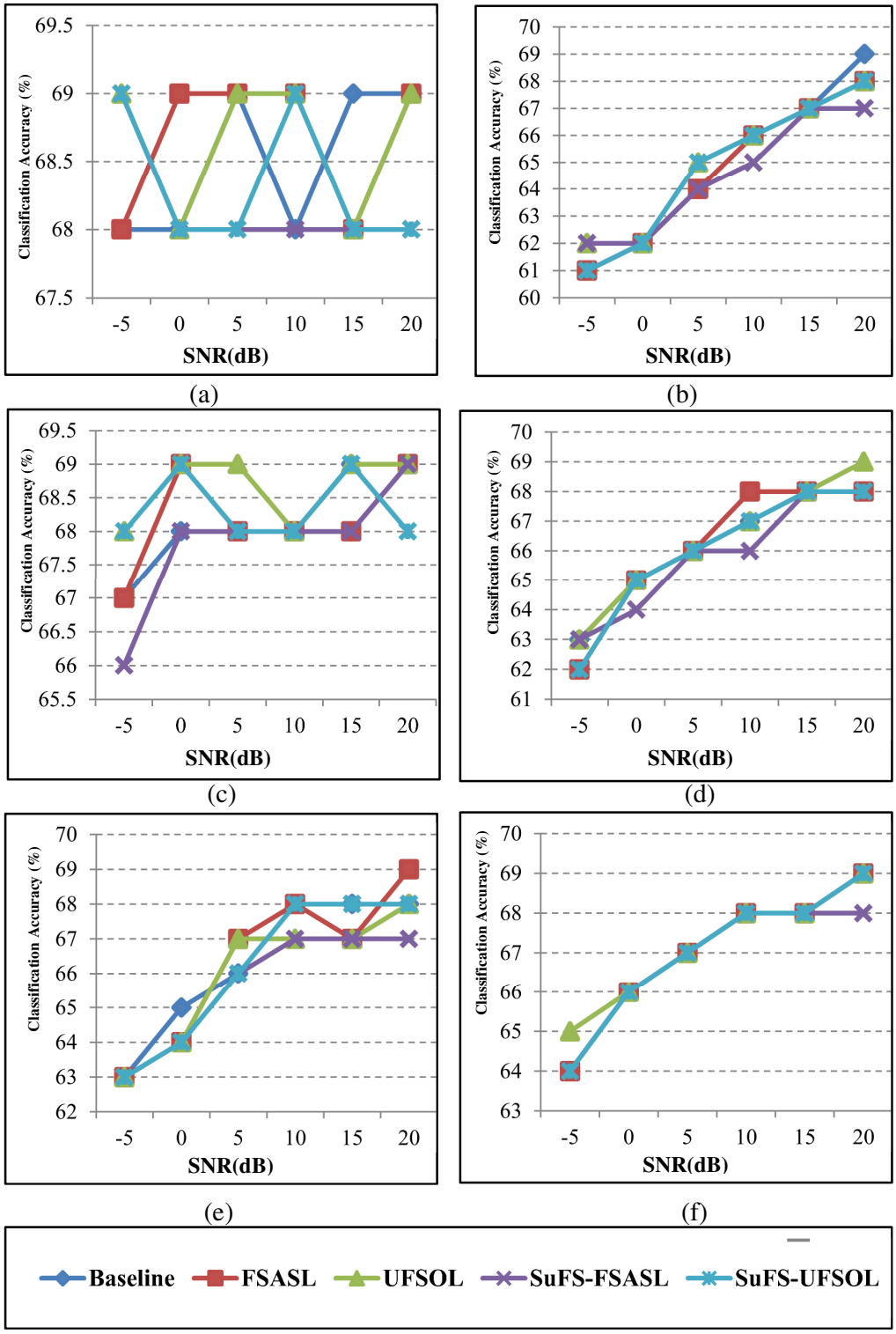


Figure 6.7: 10-Fold Cross-Validation accuracy variations of the proposed SER system for IEMOCAP noisy data with different noises (a) airport (b) babble (c) car (d) station (e) street (f) white

From figure 6.4 for EMO-DB noisy data, the proposed SER system, the testing accuracies for noisy data above 15dB are comparable with that of the clean speech database i.e., >80% using FSASL algorithm except in the case of street noise where the highest accuracy achieved for baseline and FSASL are 77% & 76% respectively. Likewise in figure 6.5, the validation accuracy is >75% for SNRs above 15dB using FSASL feature selection. For SNRs <15dB, the accuracies decreases abruptly. Similarly, for noisy IEMOCAP data, the testing accuracy of the proposed SER system is >68% for airport and car noises for the SNRs from -5dB to 20dB with the UFSOL algorithm as shown in figure 6.6. These results are comparable with that of a clean database. For other noises considered, the accuracies are >68% for SNRs above 10dB. From figure 6.7, the validation accuracy for all the noises the accuracy is >68% for all SNR levels with UFSOL except in the case of babble noise.

The results of the proposed SER system in presence of noisy conditions outperform the existing works. The comparison of the proposed SER system for EMO-DB using the FSASL algorithm in presence of white Gaussian noise is shown in table 6.4. For the remaining noises considered from the Aurora database, the proposed SER system with FSASL, compared with existing work [109] for both EMO-DB and IEMOCAP databases is shown in table 6.5.

Table 6.4: Comparison of the proposed SER system using FSASL with existing works in presence of white Gaussian noise for EMO-DB database

	-5dB	**0dB**	**5dB**	**10dB**	**15dB**	**20dB**
Ashish Tiwari et.al. (2010) [14]	-	-	37.5%	44.9%	55%	-
M Bashirpour et.al. (2016) [106]	25%	39%	54%	59%	64%	64.7%
Yongming Huang et.al. (2017) [108]	-	-	38%	50%	57%	65%
Proposed SER with FSASL	**70(±1.3)%**	**72(±1.1)%**	**76(±0.6)%**	**74(±1.0)%**	**73(±1.4)%**	**74(±1.0)%**

Table 6.5: Comparison of the proposed SER system using FSASL with existing work with noises of Aurora database for EMO-DB and IEMOCAP databases

		Classification Accuracy (%)											
		EMO-DB						**IEMOCAP**					
		-5dB	**0dB**	**5dB**	**10dB**	**15dB**	**20dB**	**-5dB**	**0dB**	**5dB**	**10dB**	**15dB**	**20dB**
Sara Sekkate et.al. (2019) [109]	Airport	-	45.33	59.24	69.16	74.27	78.29	-	37.41	38.54	39.44	40.16	40.76
	Babble	-	52.36	63.36	66.64	70.97	73.28	-	36.06	36.93	37.97	38.76	39.56
	Car	-	49.41	59.53	64.7	68.68	73.43	-	37.15	37.18	37.99	38.78	39.43
	Station	-	53.55	62.2	69.59	74.31	76.72	-	37.22	38.45	39.77	40.43	40.97
	Street	-	52.6	62.72	69.69	72.98	75.56	-	37.44	37.98	39.04	39.89	40.57
Proposed SER with FSASL	Airport	76(±0.2)	79(±0.7)	78(±0.7)	77(±1.0)	80(±0.4)	80(±1.2)	68(±0.6)	69(±0.2)	69(±1.0)	69(±0.8)	68(±0.5)	69(±1.1)
	Babble	51(±1.0)	64(±0.5)	67(±0.2)	74(±0.6)	70(±1.1)	77(±0.5)	61(±0.4)	62(±0.8)	64(±0.5)	66(±1.0)	67(±0.7)	68(±0.3)
	Car	73(±0.4)	74(±1.2)	73(±0.2)	71(±0.8)	73(±0.6)	78(±1.0)	67(±0.2)	69(±0.9)	68(±0.5)	68(±1.5)	68(±1.2)	69(±0.9)
	Station	50(±0.4)	71(±0.3)	70(±1.2)	72(±0.5)	76(±0.1)	75(±1.5)	62(±1.2)	65(±0.8)	66(±0.7)	68(±0.8)	68(±0.5)	68(±1.1)
	Street	68(±0.5)	71(±0.9)	72(±0.5)	73(±1.1)	75(±0.7)	76(±1.2)	63(±0.8)	64(±0.5)	67(±0.6)	68(±1.0)	67(±0.7)	69(±1.2)

Even though the proposed SER system is robust to noisy conditions compared to many of the existing SER works, still there is a need to further improve the robustness of the SER system in noisy conditions that has similar performance in cleaned ones. To further improve and make the SER system robust to noisy conditions, a speech de-noising method can be employed before performing the SER task. In this work, denseNMF speech de-noising is used to improve SER performance. Figures 6.8 to 6.11 show the performance of the proposed SER after de-noising using denseNMF at different noisy conditions and SNR levels from -5dB to 20dB. The same set of features selected from different feature selection algorithms considered in this work is used in the performance analysis of the de-noised SER system.

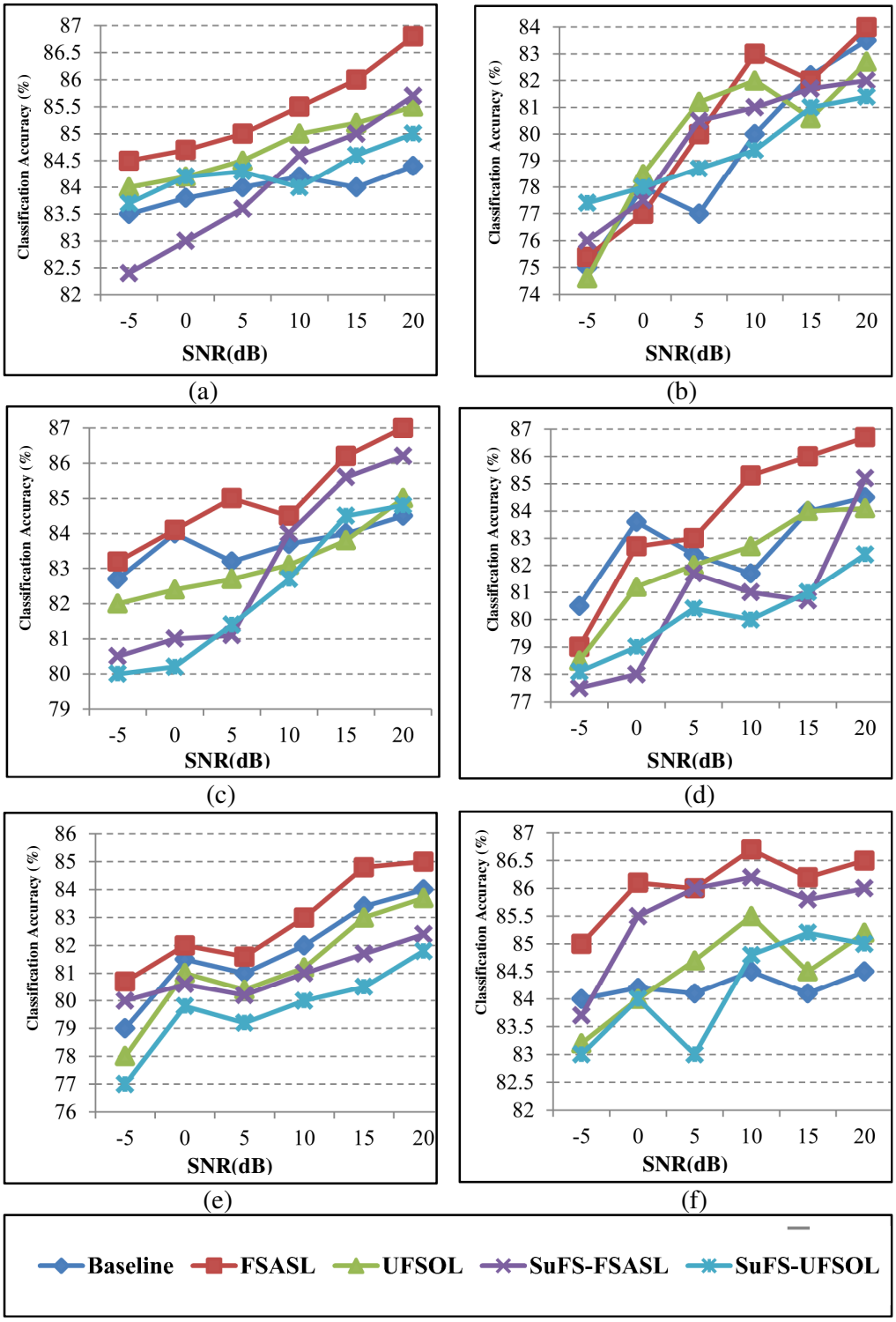


Figure 6.8: Hold-out validation accuracy variations of the proposed SER system for EMO-DB after de-noising at different noisy conditions (a) airport (b) babble (c) car (d) station (e) street (f) white

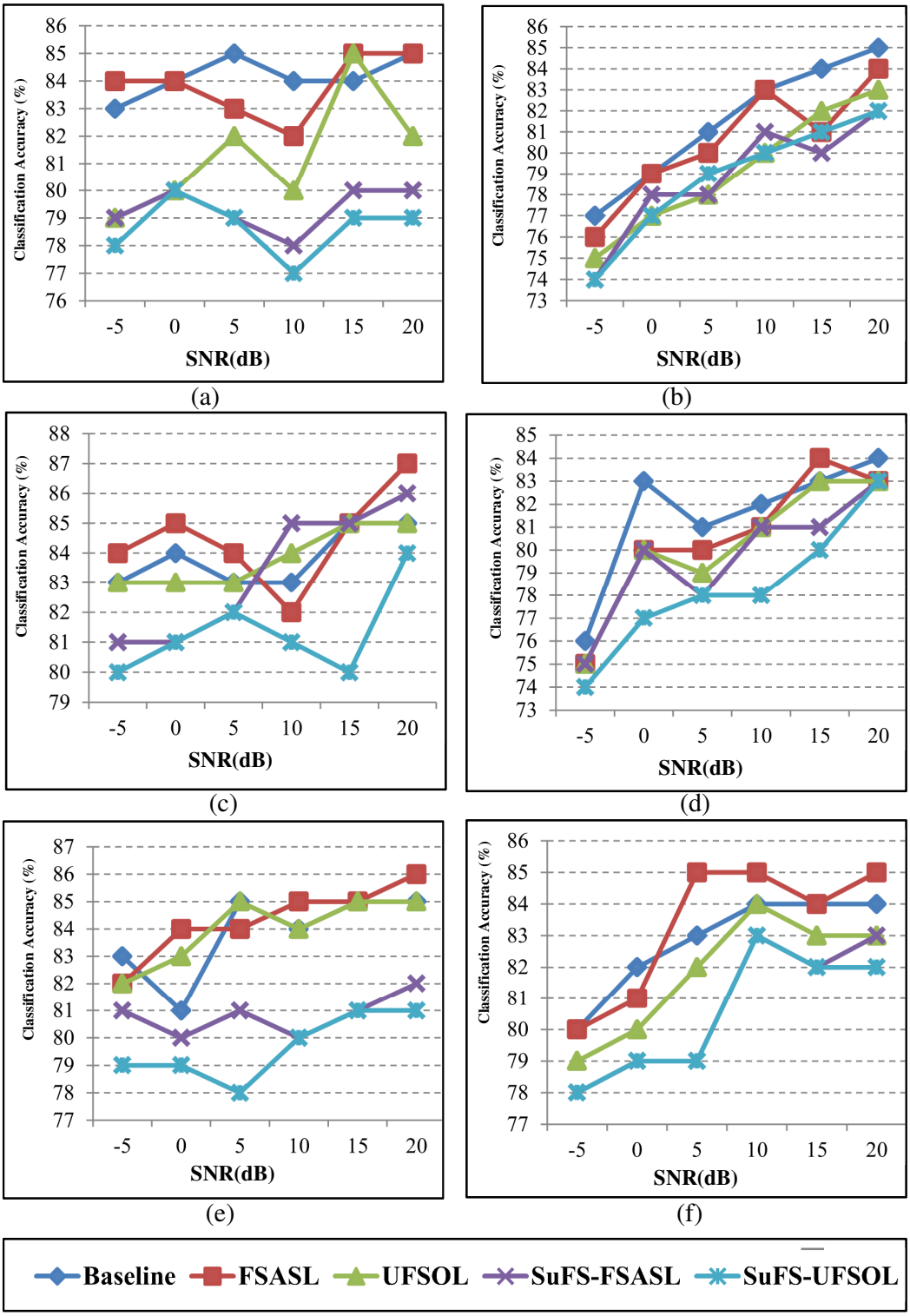


Figure 6.9: 10-fold cross-validation accuracy variations of the proposed SER system for EMO-DB after de-noising at different noisy conditions (a) airport (b) babble (c) car (d) station (e) street (f) white

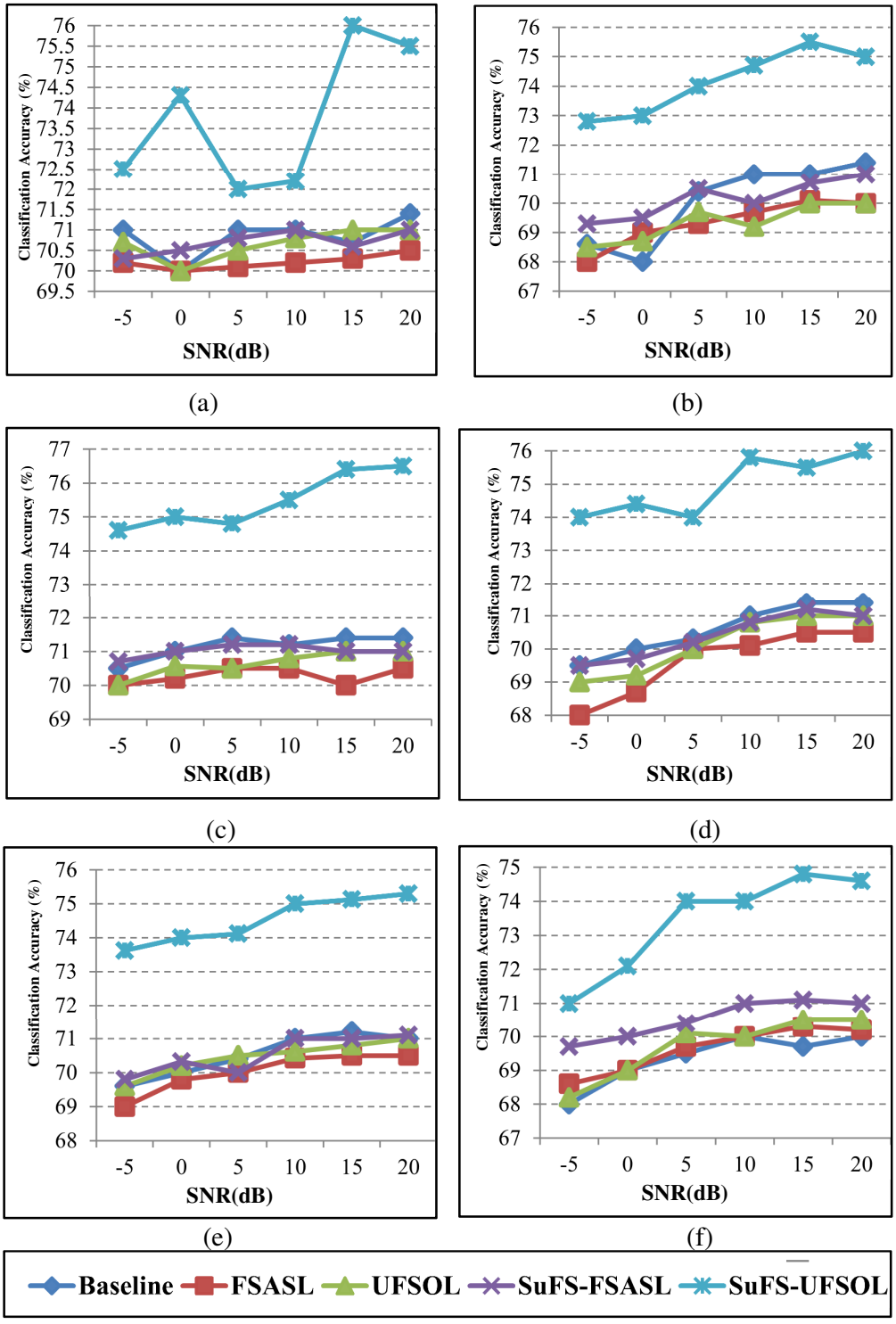


Figure 6.10: Hold-out validation accuracy variations of the proposed SER system for IEMOCAP after de-noising at different noisy conditions (a) airport (b) babble (c) car (d) station (e) street (f) white

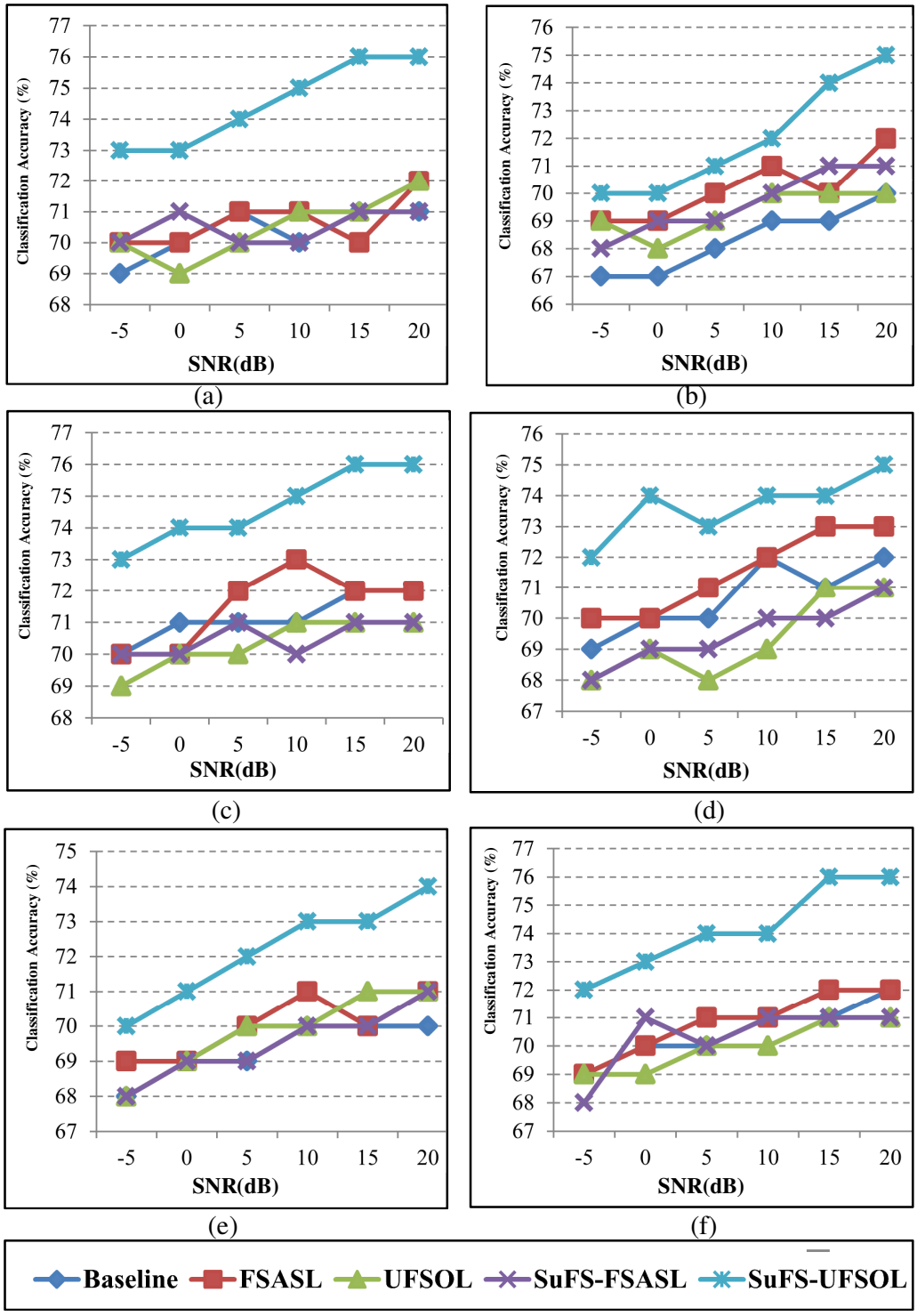


Figure 6.11: 10-Fold Cross-Validation accuracy variations of the proposed SER system for IEMOCAP after de-noising at different noisy conditions (a) airport (b) babble (c) car (d) station (e) street (f) white

Figures 6.8 and 6.9 show the hold-out and 10-fold cross-validation results of the proposed SER system after de-noising with denseNMF in terms of classification accuracy performance metric for the EMO-DB database. The FSASL based SER system performs nearly equal to the clean SER system for all the noisy conditions except in the case of babble noise for SNR levels <5dB and negative SNR level for both babble as well as station noises with 10-fold cross-validation. Whereas, UFSOL-SuFS based SER system gives better performance than the rest for babble noise in SNRs <5dB condition. Similarly, for the IEMOCAP database, figures 6.10 and 6.11 show that the UFSOL-SuFS based SER system gives better performance with both hold-out and cross-validation schemes for all noisy conditions, SNR levels considered. The accuracy, in this case, is nearly equal to the clean environment SER system. Thus, by adopting the de-noising before SER, the robustness of the proposed SER system is improved.

6.4 Summary

In the proposed work, the unsupervised feature selection algorithms are adopted for overcoming the drawbacks of the curse of dimensionality in SER and is validated both in clean and noisy conditions for robustness in real-time environments. The FSASL and UFSOL techniques are applied on the large feature set (1615 features) comprising of 1582 INTERSPEECH 2010 paralinguistic, 20 GTCC and 13 PNCC features. The EMO-DB and IEMOCAP databases are considered for the proposed SER analysis. The SVM classifier using linear and RBF kernels with 10-fold cross-validation and hold-out validation is used for emotion classification with classification accuracy and computational time as the performance metrics. For EMO-DB, the highest accuracy is achieved with the best 601 features selected using UFSOL and 504 features using the FSASL algorithm. While for IEMOCAP the highest accuracy is obtained with 1254 best features selected using both UFSOL and FSASL respectively. To reduce further the feature dimension without decreasing the classification accuracy of SER, a novel SuFS algorithm is proposed. This SuFS algorithm is applied to the optimal feature sets obtained using UFSOL and FSASL algorithms. For EMO-DB, the highest accuracy of 85($\pm$0.5)% is achieved with 454 features using UFSOL-SuFS and 85($\pm$1.0)%

with 352 features using FSASL-SuFS. While for IEMOCAP 77(±1.4)% accuracy is achieved with 802 features using UFSOL-SuFS and 71(±1.2)% with 653 features using FSASL-SuFS algorithms in the proposed SER system. To validate the robustness in noisy environments, the airport, babble, car, station and street noises of Aurora noise database along with the white Gaussian noise with SNR levels from -5dB to 20dB are considered for testing the proposed SER performance. The results clearly show that the proposed SER system outperforms the baseline and many of the existing works both in clean and noisy environments. Whereas, the proposed SER system in presence of noisy conditions at SNR levels greater than 15dB performs the same as the clean environment one. But, for SNRs less than 15dB the accuracy is reduced. To overcome this drawback, a denseNMF speech de-noising technique is adopted before SER for noise removal. It is evident that by using denseNMF, the accuracy of the noisy SER system is improved despite the system without denseNMF.

Chapter 7

Conclusions and Future Scope

This chapter gives an insight into the thesis obtained from the contributions made towards the development of speech emotion recognition system using machine learning techniques, overcoming the issues addressed in earlier chapters. The future scope is discussed with some of the potential areas of advancements in the research field of speech emotion recognition.

7.1 Conclusions

In this thesis work, four contributions have been proposed. In the first contribution in chapter 3, a speech emotion recognition system has been proposed to detect the stressed emotions. The proposed SER system has been developed using spectral and Teager energy feature fusion. Higher accuracy has been achieved for stressed emotions, by combining TEO with spectral features in the proposed system. The results shown in tables 3.1 to 3.5 and figures 3.7 to 3.9 indicate that the performance of the proposed SER system with T-MFCC, T-LPC, T-LPCC, MFCC-RASTA-PLP and T-MFCC-RASTA-PLP features provided an

improved emotion classification rate of 93%, 88%, 92%, 92% and 94.7%-96% respectively for male and female speech data. These accuracies are high compared to the SER system using MFCC (52% & 86%), Pitch (60% & 65.3%), LPC (74.6% &76%), LPC+Pitch (76% & 78.7%) and LPCC (89.3% & 92%) for EMO-DB database.

In the second contribution, a Semi-NMF based speech emotion recognition system has been developed in chapter 4. In the proposed system, the Semi-NMF technique with SVD initialization has been employed to optimize MFCC, LPCC and TEO-AutoCorr feature sets. The performance of the proposed SER system has been analyzed with EMO-DB and IEMOCAP databases using k-NN and SVM classifiers. A 5-fold Cross-validation scheme has been used to train the feature sets to consider the entire dataset for both training and testing, to avoid overfitting problem. The combination of the optimized feature sets of MFCC, LPCC and TEO-AutoCorr have been used in the proposed SER system. The results depicted in figures 4.5 to 4.8 and tables 4.2, 4.3 prove that the highest classification accuracies have been achieved by the proposed system with the optimized feature set. Using SVM and k-NN classifiers, the accuracies are 90.12% and 89.3% for the EMO-DB database and for the IEMOCAP database, they are 83% and 78% respectively.

The semi-NMF algorithm used for feature optimization is a transformation technique and it lacks data interpretability. So, an SER system with UFSOL, FSASL and SuFS unsupervised feature selection algorithms has been proposed for feature optimization with data interpretability in the third contribution in chapter 5. The INTERSPEECH 2010 Paralinguistic and GTCC features have been optimized using feature selection algorithms in the proposed system. From the results shown in tables 5.3 and 5.4, it is observed that the highest accuracy achieved by the proposed SER system with FSASL-SuFS for EMO-DB using hold-out validation is 86% with a training time of 0.165 seconds, testing time of 0.0.32 seconds, and using 10-fold cross-validation the accuracy is 85% with a computation time of 1.4 seconds. For the IEMOCAP database, the highest accuracy achieved for the proposed SER system with UFSOL-SuFS using hold-out validation is 77.5% with a training time of 14

seconds, testing time of 2.9 seconds, and using 10-fold cross-validation the accuracy is 77% with a computation time of 125.5 seconds.

The SER systems proposed in chapters 3 to 5 have been developed with a clean speech database. Hence, for the SER system not to be susceptible to noisy conditions, in the fourth contribution, a noise robust SER system using PNCC features and denseNMF de-noising has been proposed in chapter 6. The combination of INTERSPEECH 2010 paralinguistic set, GTCC and PNCC features have been used in the proposed system to achieve noise robustness. The unsupervised feature selection algorithms proposed in chapter 5 have been used to optimize this huge feature set. Initially, the SER system is developed without applying de-noising and the results are shown in figures 6.4 to 6.7. For the SNR levels above 15dB with different noises, the PNCC based SER system with FSASL performs similar to the system developed in a clean speech environment in terms of accuracy. To further improve SER accuracy at lower SNR levels, a denseNMF speech de-noising technique has been used before performing SER. The results of the proposed system after de-noising are shown in figures 6.8 to 6.11. For all the considered noisy conditions other than babble noise, for the speech emotion recognition system with FSASL, the accuracy is higher than 80% for EMO-DB at lower SNR levels. Whereas for the IEMOCAP database, using UFSOL-SuFS based SER system with de-noising, the accuracy is greater than 71% for all the considered noisy conditions even at lower SNR levels.

7.2 Future Scope

The speech emotion recognition system can further be upgraded to cross-corpus analysis, such that the developed SER system can be language independent. The speech signal is used as a mode of data source for emotion recognition in the proposed work. This system can be further improved to a multimodal emotion recognition system, where more than one data source such as facial, speech, etc. can be used to acquire better emotion accuracy.

Appendix I

Table I: Best INTERSPEECH 2010 paralinguistic features selected using UFSOL, FSASL, UFSOL-SuFS and FSASL-SuFS algorithms for the proposed SER system for EMO-DB and IEMOCAP databases

Method	EMO-DB			IEMOCAP		
FSASL	Position – max.		For all functionals and their deltas	Position – max.		For all functionals and their deltas
	Position – min.		For all functionals and their deltas except F0 Env	Position – min.		For all functionals and their deltas
	Arithmetic mean		*F0 Sub,F0 Env, MFCC[1,3-14]*	Arithmetic mean		*F0Sub+Δ,F0Env+Δ,Voicing Prob,JitterLocal,JitterDDP,ShimmerLocal, PCM,MFCC[0-14],MFCCΔ[0,1,3,5-7,9-14],LogMel[0-7],LSP [1-7]*
	Standard Deviation		*F0 Sub, F0 Env+Δ,MFCC[1-14], MFCCΔ[0,1,3,4,7,9-14]*	Standard Deviation		*F0 Sub+Δ,F0 Env+Δ,Jitter Local,Jitter DDP, Shimmer Local, PCM, MFCC+Δ[0-14], Log Mel [0-5,7], Log Mel Δ [0-7], LSP [1-7]*
	Skewness		*F0 Sub Δ, F0 Env+Δ, Voicing Prob, Jitter Local Δ, Jitter DDP+Δ,Shimmer Local+Δ,LSP[6],LSPΔ[6,7],MFCCΔ[1]*	Skewness		For all functionals and their deltas
	Kurtosis		*F0 Sub+Δ, F0 Env+Δ,Voicing Prob+Δ, Jitter Local+ Δ, Jitter DDP+ Δ,Shimmer Local+ Δ, PCM Δ,MFCC[0,3-5,10,12,14], MFCCΔ[2,4-10,12,13], Log Mel [0-3,6], Log Mel Δ [1-7], LSP [0,5-7], LSPΔ [0-3,6]*	Kurtosis		For all functionals and their deltas
	Linear regression coefficient	c1	*F0 Sub*	Linear regression coefficient	c1	*F0 Sub+Δ,F0 Env+Δ, MFCC[0-2,4-14]*
		c2	*F0 Sub, F0 Env+Δ, MFCC[0-14]*		c2	*F0 Sub+Δ,F0 Env+Δ,Voicing Prob, Shimmer Local, PCM, MFCC+Δ[0-14],Log Mel [0-7], Log Mel Δ [0-6], LSP [0-7]*
	Linear regression error	A	*F0 by Sub, F0 Env, MFCC[0-13], MFCCΔ[1,7,9,10]*	Linear regression error	A	*F0 Sub+Δ,F0 Env+Δ, Jitter Local, Jitter DDP, Shimmer Local,PCM, MFCC+Δ[0-14], Log Mel+Δ [0-7]*
		Q	*MFCC[0-14], MFCCΔ[0-7,9-13], F0 Sub + Δ, F0 Env + Δ, Log Mel [0-6]*		Q	*F0 Sub+Δ,F0 Env+Δ, Shimmer Local, PCM, MFCC [0-14], MFCCΔ[0-13],Log Mel [1-7], Log Mel Δ [0-7]*
		1	*F0 Sub, F0 Env, Log Mel [0,1,4], MFCC[0-12,14], MFCCΔ[1,10,14]*		1	*F0 Sub+Δ,F0 Env+Δ, Voicing Prob, Shimmer Local, PCM, MFCC+Δ[0-14], Log Mel+Δ [0-7], LSP [0-7]*

	Quartile	2	*F0 Sub, F0 Env, MFCC[0-14]*	Quartile	2	*F0 SubΔ,F0 Env+Δ, Voicing Prob, Shimmer Local, MFCC+Δ[0-14], Log Mel [0-7], Log Mel Δ [0,1,4-6], LSP [1-7]*
		3	*F0 Sub, F0 Env+ Δ, MFCC[0-11,13,14], MFCCΔ[0,3,7,10,11,14], Log Mel [3-5]*		3	*F0 Sub+Δ,F0 Env+Δ, Voicing Prob, Shimmer Local, PCM, Jitter Local, Jitter DDP, MFCC+Δ[0-14], Log Mel+Δ [0-7], LSP [1-7]*
	Quartile range	2-1	*F0 Sub, F0 Env, MFCC[0-4,6-9,11-14], MFCCΔ[7,10], Log Mel [0,5]*	Quartile range	2-1	*F0 Sub+Δ,F0 Env+Δ, Voicing Prob, Shimmer Local, PCM, MFCC+Δ[0-14], Log Mel [0-7], Log Mel Δ [0,2-7], LSP [0,2,3,5]*
		3-1	*F0 Sub, F0 Env+ Δ, MFCC[0-14], MFCCΔ[0,1,3,4,6,7,9-12,14], Log Mel [0-7]*		3-1	*F0 Sub+Δ,F0 Env+Δ, Voicing Prob, Shimmer Local, PCM+Δ, Jitter Local, Jitter DDP, MFCC+Δ[0-14], Log Mel+Δ [0-7], LSP [0,3,5]*
		3-2	*F0 Sub, F0 Env+ Δ, MFCC[0-4,6-14], MFCCΔ[0,1,3,6,9,10,14], Log Mel [3-5]*		3-2	*F0Sub+Δ,F0Env+Δ,VoicingProb, ShimLocal,PCM,JitteLocal,Jitter DDP, MFCC+Δ[0-14],LogMel[0-7],LogMelΔ [0,1,3-6], LSP [3,4]*
	Percentile	99.0	*F0 Sub, F0 Env+ Δ, MFCC[0-14], MFCCΔ[1-14], Log Mel [1-7], PCM*	Percentile	99.0	*F0Sub+Δ,F0Env+Δ,VoicingProb,ShimLocal, PCM+Δ, Jitter Local, JitterDDP,MFCC[0,2-14],MFCCΔ[0-14],LogMel+Δ[0-7],LSP[0-7]*
		1.0	*F0 Env Δ, MFCC+Δ[0-14], Log Mel [3-7]*		1.0	*F0Env+Δ,VoicingProb,PCMΔ,MFCC+Δ[0-14],LogMel+Δ[0-7], LSP[0-7]*
	Percentile range		*F0 Env, MFCC+Δ[0-14], Log Mel [1,2,5,6], Log Mel Δ[0,1,7], PCM*	Percentile range		*F0 Env+Δ, VoicingProb, PCM+Δ, MFCC+Δ[0-14], LogMel[0-7], LogMelΔ [0-2,4-7], LSP [0-7], LSPΔ[2]*
	Up-level time 90		*Jitter Local Δ*	Up-level time	75	*F0 Sub+Δ,F0 Env+Δ,Voicing Prob,Jitter Local+Δ, Jitter DDP, Shimmer Local+Δ,PCMΔ, MFCC[0-3,5-10,12-14],MFCCΔ[0-14], LogMel[0-6], LogMelΔ [0,4,5,7], LSP [0-7], LSPΔ[0-2,4,6,7]*
					90	*F0Sub+Δ,F0Env,VoicingProb,JitterLocal+Δ,JitterDDP+Δ, ShimmerLocal+Δ, MFCC[0,1,5-7,9,13],MFCCΔ[2,4,7,12], LogMel[0-2,4-6], LSP [0,2,3,5-7], LSPΔ[2]*
UFSOL	Position – max.		For all functionals and their deltas except LogMelΔ[6]	Position – max.		For all functionals and there deltas
	Position – min.		For all functionals and their deltas except F0 Env	Position – min.		For all functionals and there deltas
	Arithmetic mean		*F0 Sub,F0 Env, PCM, MFCC[0-14]*	Arithmetic mean		*F0SubΔ,F0Env+Δ,VoicingProb,,JitterDDP+Δ,ShimmerLocal+Δ, PCM, MFCC[0-14],MFCCΔ[3-6,8,12,14],LogMel[0-7],LSP[0-7]*
	Standard Deviation		*F0 Sub, F0 Env, MFCC[0,2,5,8-12]*	Standard Deviation		*F0 Sub+Δ,F0 Env+Δ, Jitter Local+Δ, Jitter DDP+Δ,Shimmer Local+Δ,PCM, MFCC+Δ[0-14], Log Mel+Δ [0-7], LSP[0-2,5]*
	Skewness		*F0 Sub Δ, Jitter Local Δ, Jitter DDP+Δ,Shimmer LocalΔ,*	Skewness		For all functionals and their deltas
	Kurtosis		*F0 Sub+Δ,F0 Env+Δ,Voicing Prob,Jitter Local+Δ, Jitter DDP+Δ,Shimmer Local+Δ,PCM, Log Mel [6], Log Mel Δ [2,4,6,7], LSP[6,7], LSPΔ [4,6]*	Kurtosis		For all functionals and their deltas
	Linear regression coefficient	c1	*PCM*	Linear regression coefficient	c1	*F0 Sub+Δ,F0 Env+Δ, PCM, MFCC[0-11,13,14]*
		c2	*F0 Sub,F0 Env+Δ,PCM,MFCC[0-14],Log Mel[1,2,4]*		c2	*F0Sub+Δ,F0Env+Δ,JitterLocal,JitterDDP,ShimmerLocal+Δ,PCM, VoicingProb,MFCC+Δ[0-14],LogMel[0-7],LSP[0-7],LogMelΔ[0-2]*
	Linear regression error	A	*F0 Sub,F0 Env, PCM, MFCC[0-4,7,8,10]*	Linear regression error	A	*F0 Sub+Δ,F0 Env+Δ, PCM, Shimmer Local+Δ, MFCC+Δ[0-14], Log Mel+Δ [0-7], LSP[0]*
		Q	*F0 Sub,F0 Env+Δ,PCM,MFCC+Δ[0-14],Log Mel[0-7]*		Q	*F0Sub+Δ,F0Env+Δ,PCM,JitterDDPΔ,MFCC+Δ[0-14],LogMel[0-7], LogMelΔ[1-7]*
	Quartile	1	*F0 Sub,F0 Env,MFCC[0-14],Log Mel[0,2,3,5,6]*	Quartile	1	*F0Sub+Δ,F0Env+Δ,Jitter DDP,Shimmer Local+Δ, PCM,Voicing Prob,MFCC+Δ[0-14], Log Mel+Δ [0-7], LSP [0-7]*
		2	*F0 Sub,F0 Env,MFCC[0-14],Log Mel[6,7]*		2	*F0Sub+Δ,F0Env+Δ,VoicingProb, Jitter DDP,Shimmer Local+Δ, PCM,JitterLocal, MFCC+Δ[0-14],Log Mel [0-7], LSP [0-7]*
		3	*F0 Sub,F0 Env,MFCC[0-14]*		3	*F0Sub+Δ,F0Env+Δ,VoicingProb, Jitter DDP+Δ,Shimmer Local, PCM,JitterLocal, MFCC+Δ[0-14],Log Mel +Δ [0-7], LSP [0-7]*
		2-1	*F0 Sub,F0 Env,MFCC[0-7,9,10,12-14]*	Quartile	2-1	*F0Sub+Δ,F0Env,ShimmLocal+Δ,PCM,MFCC+Δ[0-14],LogMel+Δ[0-7],LSP[6]*

	Quartile range	3-1	*F0 Sub,F0 Env+Δ,MFCC[0-14], MFCCΔ[0,1,2,6,7], Log Mel[1]*	range	3-1	*F0Sub+Δ,F0Env+Δ,Shimmer Local+Δ, PCM, VoicingProb Δ, Jitter DDPΔ, MFCC+Δ[0-14],Log Mel +Δ [0-7], LSP [0-3,5]*
		3-2	*F0 Sub,F0 Env, MFCC[0-3,5,6,8,9,13,14]*		3-2	*F0Sub+Δ,F0Env+Δ,Shimmer Local+Δ, PCM, MFCC+Δ[0-14], Log Mel [0-7], Log Mel Δ [1-7], LSP [1,2]*
	Percentile	99.0	*F0 Sub,F0 Env+Δ,MFCC[0-14], MFCCΔ[0-5,7-12],Log Mel[2,3]*	Percentile	99.0	*F0Sub+Δ,F0Env+Δ,Shimmer Local+Δ, Jitter DDP+Δ, PCMΔ, LSPΔ[1], , LSP [0-7], VoicingProb, Jitter Local, MFCC+Δ[0-14],Log Mel +Δ [0-7]*
		1.0	*F0 Env+Δ,MFCC[0-14], MFCCΔ[0-4,6-12],Log Mel[0,2,3,5-7]*		1.0	*F0Env+Δ, PCM+Δ, VoicingProb, MFCC+Δ[0-14],Log Mel +Δ [0-7], LSP [0-7], LSPΔ[0,1]*
	Percentile range		*F0Env+Δ,MFCC+Δ[0-14],LogMel[0,2,3,5-7], LogMelΔ[2]*	Percentile range		*F0Env+Δ, PCM+Δ, VoicingProb+Δ, MFCC+Δ[0-14],Log Mel +Δ [0-7], LSP [0-7], LSPΔ[1-6]*
	Up-level time 75		*Shimmer Local Δ*	Up-level time	75	*F0Sub+Δ,F0Env,Shimmer Local+Δ, Jitter DDP+Δ, VoicingProb, Jitter Local+Δ, MFCC+Δ[0-14], Log Mel[0-7],Log MelΔ[2,3,6, 7], LSP[0-7], LSPΔ[0-3,5-7]*
					90	*F0Sub+Δ,F0Env,ShimmerLocal+Δ,JitterDDP+Δ,VoicingProb,JitterLocal+Δ,MFCC[1,6,13,14],MFCCΔ[2],LogMel[0,1,6,7],LSP[1,3-7], LSPΔ[5]*
FSASL-SuFS	Position – max.		*F0 Sub+Δ,F0 Env,Voicing Prob+Δ,Jitter Local+Δ, Jitter DDP+Δ,Shimmer Local,PCM+Δ, MFCC[0,2,3,5-8,10,11], MFCCΔ[0-3,9,11,14],Log Mel[0-2,6,7],Log MelΔ[0-2,4,6,7], LSP[1,2,5,7], LSPΔ[0,3-7]*	Position – max.		*F0SubΔ,F0Env,Shimmer Local+Δ, Jitter DDP+Δ, VoicingProb,PCM+Δ, MFCC[0,3-14], MFCCΔ[0,2-6,8-14], Log Mel[0-3,5-7], Log MelΔ[0-3,5,7], LSP[0,4-7], LSPΔ[0-3,5-7]*
	Position – min.		*F0 Sub+Δ,F0 Env Δ,Voicing Prob+Δ, Jitter DDP+Δ, MFCC[1-3,5,7,11,12], MFCCΔ[0-5,7,9,10,12-14],Log Mel[0,1,6,7],LogMelΔ[1,3-7], LSP[2,3,5-7], LSPΔ[2-7]*	Position – min.		*F0Sub+Δ,F0Env+Δ,Shimmer Local, Jitter DDP+Δ, VoicingProb+Δ, Jitter LocalΔ,PCM, MFCC[3-5,7,9,10,12-14], MFCCΔ[0,1,5-14], Log Mel +Δ [0-7], LSP[0,2-7], LSPΔ[0-7]*
	Arithmetic mean		*MFCC[1,3,4,8,10-13]*	Arithmetic mean		*F0SubΔ,Voicing Prob, JitterLocal, JitterDDP, PCM,MFCC[0,2-9,11-13], MFCCΔ[0,1,3,6,7,11-14],LogMel[1,4],LSP [2]*
	Standard Deviation		*F0 Env Δ, MFCC[2-4,6,7,9,10,12,13], MFCCΔ[7,10]*	Standard Deviation		*Jitter Local, PCM, MFCC[2,3,9,11], MFCCΔ[3,11,14], Log Mel [1,2,6], Log Mel Δ [2,5], LSP [0,2,3]*
	Skewness		*F0SubΔ,F0EnvΔ,Voicing Prob,Jitter Local Δ,Shimmer LocalΔ*	Skewness		*Shimmer Local+Δ,Jitter Local,PCM, MFCC[10,12], MFCCΔ[1,3,5,7 Log Mel [3,6,7], Log Mel Δ [0,4], LSP [1,6],LSPΔ[7]*
	Kurtosis		*F0SubΔ,F0Env,VoicingProb+Δ,JitterLocal,Shimmer LocalΔ, Jitter DDP+Δ, MFCC[0,3,12,14], MFCCΔ[2,9,10,12], Log Mel[0,1,2],Log MelΔ[4,6,7], LSP[0,6,7], LSPΔ[7]*	Kurtosis		*F0SubΔ,F0Env,Shimmer LocalΔ, Jitter DDP+Δ, VoicingProbΔ, Jitter Local, PCM, MFCC[5,9,10,13], MFCCΔ[1,3,7-9,11], Log Mel [0-3,6,7], Log Mel Δ [0,3,5-7], LSP [3-5],LSPΔ[0-2,5-7]*
	Linear regression coefficient	c1	*F0 Sub*	Linear regression coefficient	c1	*F0 SubΔ,F0 Env+Δ, MFCC[1,2,5-8,10-13]*
		c2	*F0 Sub, F0 Env+Δ, MFCC[0,2-4,6,8,10,12]*		c2	*F0 SubΔ, Voicing Prob, MFCC[0,2-8,11,12-14],MFCCΔ[1,2,12,13],Log Mel [0,1,4,5], Log Mel Δ [0,2,4-6], LSP [1,7]*
	Linear regression error	A	*F0 Sub, F0 Env, MFCC[1-8,10,11,13]*	Linear regression error	A	*F0 EnvΔ, Jitter Local, Jitter DDP, PCM, MFCC[0-2,4,5,7-9,12-14], MFCCΔ[1,3,6,12], Log Mel[1,2,5],Log MelΔ [1,2,5,7]*
		Q	*MFCC[0,2,5,7-14], MFCCΔ[0-5,9,12-13], F0 Sub, F0 Env, Log Mel [0-6]*		Q	*F0 EnvΔ, Shimmer Local, PCM, MFCC [1,3,6,7,11,13,14], MFCCΔ[0-2,4-8,10,12],Log Mel [0-5], Log Mel Δ [0,2,6,7]*
	Quartile	1	*F0 Sub, Log Mel [0], MFCC[0,1,4,6,9-11,14], MFCCΔ[10]*	Quartile	1	*F0 Sub,Voicing Prob,MFCC[1,3-7,10-14],MFCCΔ[1,3,4,6,10,11], Log Mel[1,6],LogMelΔ [0,2-4], LSP [0-2]*
		2	*MFCC[0-4,6-9,11-13]*		2	*F0 Sub,F0 EnvΔ, Voicing Prob, Shimmer Local, MFCC[1,2,4-11,13,14], MFCCΔ[0,2,3,5,7,8,11], Log Mel [0,1,3,4], Log Mel Δ [0,1,4,5], LSP [2]*
		3	*F0 Env+ Δ, MFCC[0-2,4,6-11,13], Log Mel [4]*		3	*F0 EnvΔ, Jitter Local, Jitter DDP, MFCC[0,2-4,8-10,13,14], MFCCΔ[3,8,12,13] Log Mel[0-3,6,7]LogMelΔ [1-3], LSP [1,2,7]*

	Quartile range	2-1	*F0 Sub, F0 Env, MFCC[2,4,6,8,9,11,14], MFCCΔ[10], Log Mel [0]*	Quartile range	2-1	*F0 Sub,F0 Env+Δ, Voicing Prob, Shimmer Local, PCM, MFCC[0,1,5-9,12], MFCCΔ[0-2,7,13], Log Mel[1,2],LogMelΔ [0,2,3,5], LSP [0,2,3,5]*
		3-1	*MFCC[3,6,10,11,13,14], MFCCΔ[0,1,3,4,6,7,9,10,11,14], Log Mel [0,3-6]*		3-1	*F0 SubΔ, Voicing Prob, PCM,Jitter Local, Jitter DDP, MFCC[0-4,6-8,10-12,14],MFCCΔ[2,5,10,15], Log Mel[0-3,6],LogMelΔ [0,3,7], LSP [3,5]*
		3-2	*F0 Env, MFCC[0,2,4,7,8,10,11-14], Log Mel [4]*		3-2	*F0SubΔ,F0EnvΔ,VoicingProb, PCM, JitterLocal, Jitter DDP, MFCC[2-13], LogMel[0,2], LogMelΔ [1,4,5], LSP [3,4]*
	Percentile	99.0	*MFCC[1-6,9,11], MFCCΔ[2-7,9-13], Log Mel [1-3,6,7]*	Percentile	99.0	*F0SubΔ, VoicingProb,ShimLocal, PCM+Δ, MFCC[4-6,8-11,13], MFCCΔ[0,1,3,5,8,9,13], LogMel[2],LogMelΔ[0,2,6,7],LSP[0,2,6,7]*
		1.0	*F0 Env Δ, MFCC+Δ[0-14], Log Mel [3-7]*		1.0	*F0 Env+Δ, VoicingProb, PCMΔ, MFCC[1-3,6-8,10-14],MFCCΔ[1-5,7-9,11,13,14], LogMel[4,7], LogMel Δ[1,3,7], LSP[1,4]*
	Percentile range		*MFCC[3,7-9],MFCCΔ[0-2,4,6-9,12,13],Log Mel [1,2,5], Log Mel Δ[0]*	Percentile range		*VoicingProb, MFCC[2,4-6,8,9,12,13],MFCCΔ[0,2,6-9,11,13], LogMel[3,5], LogMelΔ [2,5,6], LSP [0,1,3,7], LSPΔ[2]*
	Up-level time		—	Up-level time	75	*F0 EnvΔ, Jitter DDP, MFCC[3,5,8,12,14],MFCCΔ[0,2,8,9,11,13,14], LogMel[1,2,4,5], LogMelΔ [0,4,5,7], LSP [0-3,5], LSPΔ[0,4,7]*
					90	*Voicing Prob, Jitter DDP, Shimmer LocalΔ, MFCC[5-7,9,13], MFCCΔ[4,7,12], LogMel[0,2,4-6], LSP [0,2,3,5-7], LSPΔ[2]*
UFSOL -SuFS	Position – max.		*F0 Sub+Δ,F0 Env+Δ,Voicing Prob+Δ,Jitter Local+Δ, Jitter DDP+Δ,Shimmer Local, PCM, MFCC[0-7,10-14], MFCCΔ[0-3,5-13], Log Mel [0,5-7], Log Mel Δ[4-7], LSP [0-6], LSPΔ [0,2,3-7]*	Position – max.		*F0 SubΔ,F0 Env,Voicing Prob+Δ,Jitter Local,,Shimmer LocalΔ, PCMΔ, MFCC[3-9,12], MFCCΔ[1,3,5,6,8-10], Log Mel [0,1,3,4,7], Log Mel Δ[0,3,5], LSP [0,2,7], LSPΔ [1,3-7]*
	Position – min.		*F0 Sub+Δ,Voicing Prob+Δ, Jitter DDP+Δ,Shimmer Local+Δ, PCM, MFCC[0-3,7-10,14], MFCCΔ[3-13], LogMel[0,1,4-7],LogMel Δ[3-7],LSP[3-5,7],LSPΔ[3-7]*	Position – min.		*F0 SubΔ,F0 EnvΔ,Voicing Prob,Jitter LocalΔ, Jitter DDPΔ,Shimmer LocalΔ, PCM+Δ, MFCC[1-3,6,7,12,13], MFCCΔ[0,3,5-8,10,12,13], Log Mel [4,7], Log Mel Δ[3,5], LSP [1,3,4,6,7], LSPΔ [2,4-6]*
	Arithmetic mean		*F0 Sub,F0 Env, PCM, MFCC[0-14]*	Arithmetic mean		*F0SubΔ,F0Env+Δ,VoicingProb,,JitterDDPΔ,ShimmerLocalΔ,MFCC[1,3,5,6,8, 11,12], MFCCΔ[3-6,8,12,14],LogMel[2,5,6],LSP[0,1,4-7]*
	Standard Deviation		*F0 Sub, F0 Env, MFCC[0,2,5,8-12]*	Standard Deviation		*F0Sub+Δ,F0EnvΔ,JitterLocal,JitterDDP+Δ,ShimmerLocal+Δ,PCM, MFCC[1,2,6,7,11-14],MFCCΔ[0,3,4,8,12],LogMel[1-3,6],LogMelΔ[0-7],LSP[0-2,5]*
	Skewness		*F0 Sub Δ, Jitter Local Δ, Jitter DDP+Δ,Shimmer LocalΔ*	Skewness		*F0 Sub+Δ,F0 Env+Δ,Voicing ProbΔ,Jitter Local+Δ, Jitter DDP+Δ, Shimmer Local+Δ, PCM+Δ, MFCC[0,,3-5,8,10,11,13,14], MFCCΔ[0,5,7-9,14], Log Mel [0-3,5,7], Log Mel Δ[0,3-7], LSP [0,1,3-7], LSPΔ [2,3,6]*
	Kurtosis		*F0 Sub+Δ,F0 Env+Δ,Voicing Prob,Jitter Local+Δ, Jitter DDP+Δ,Shimmer Local+Δ,PCM, Log Mel [6], Log Mel Δ [2,4,6,7], LSP[6,7], LSPΔ [4,6]*	Kurtosis		*F0Sub,F0 Env+Δ, Jitter LocalΔ,Shimmer LocalΔ, PCM+Δ, MFCC[1,3-5,8,12-14], MFCCΔ[0,3,7,9,10,13,14], Log Mel [1-5], Log Mel Δ[0-3,6], LSP [0,1,3,4,6,7], LSPΔ [0,5-7]*
	Linear regression coefficient	c1	*PCM*	Linear regression coefficient	c1	*F0 Sub,F0 EnvΔ,MFCC[0-11,13,14]*
		c2	*F0 Sub,F0 Env+Δ,PCM,MFCC[0-14],Log Mel[1,2,4]*		c2	*F0Env,JitterLocal,JitterDDP,ShimmerLocal+Δ,VoicingProb,MFCC[3,4,6,13,14],MFCCΔ[2,4-6,8],LogMel[2,4,5],LSP[0,3-7],LogMelΔ[0-2]*
	Linear regression error	A	*F0 Sub,F0 Env, PCM, MFCC[0-4,7,8,10]*	Linear regression error	A	*F0 Sub+Δ,F0 Env+Δ, Shimmer Local+Δ, MFCC[0,2-5,7,8,10,12-14], MFCCΔ[1,5,6,7,9,10,13,14], LogMel[0,3,4,7],LogMelΔ [0-7], LSP[0]*
		Q	*F0 Sub,F0 Env+Δ,PCM,MFCC[0,3-7,12,13],MFCCΔ[0,3-12],Log Mel[0-5,7]*		Q	*F0 SubΔ,F0 Env+Δ,Jitter DDP Δ, MFCC[0-14],MFCCΔ[0,3-9,11], Log Mel [1-4,6], Log Mel Δ [1,3-7]*
	Quartile	1	*F0 Sub,F0 Env,MFCC[0-14],Log Mel[0,2,3,5,6]*	Quartile	1	*F0SubΔ,JitterDDP,ShimmerLocal,PCM,VoicingProb,MFCC[0,6,8,11,12,14], MFCCΔ[0,1,6,12-14], LogMel[2,4,5,7],LogMelΔ [0-7], LSP [0,1,4-7]*
		2	*F0 Sub,F0 Env,MFCC[0-14],Log Mel[6,7]*		2	*F0EnvΔ,VoicingProb, Jitter DDP,Shimmer Local,JitterLocal,*

						MFCC[0,1,4-6,8,10,12,13],MFCCΔ[0-9,11-14],Log Mel [2,6], LSP [0,1,4-7]
		3	*F0 Sub,F0 Env,MFCC[0-14]*		3	*F0Sub+Δ,VoicingProb,JitterDDP+Δ,ShimmLocal,PCM,JitterLocal, MFCC[0,4,6-8,10-12],MFCCΔ[0-2,7-9,11-14],LogMel[1,6,7], LogMelΔ[0-7],LSP[0,1,3-7]*
	Quartile range	2-1	*F0 Sub,F0 Env,MFCC[0-7,9,10,12-14]*	Quartile range	2-1	*F0SubΔ,Shimmer Local+Δ,PCM,MFCC[1-5,8,10,12],MFCCΔ[0,4,7-9,11-14], LogMel[0,1,5,6], LogMelΔ[0-7],LSP [6]*
		3-1	*F0 Sub,F0 Env+Δ,MFCC[0-14], MFCCΔ[0,1,2,6,7],Log Mel[1]*		3-1	*F0SubΔ,Shimmer LocalΔ, VoicingProb Δ, Jitter DDPΔ, MFCC[1,4,5,8-14],MFCCΔ[0,1,10-14],LogMel[1,4,6], LogMel [0-7], LSP [0-3,5]*
		3-2	*F0 Sub,F0 Env,MFCC[0-3,5,6,8,9,13,14]*		3-2	*F0Sub,F0EnvΔ,Shimmer LocalΔ, PCM, MFCC[1,2,8,9,11,12,14], MFCCΔ[1,2,7-9,11-14], LogMel [2-7], Log Mel Δ [1-7], LSP [1,2]*
	Percentile	99.0	*F0 Sub,F0 Env+Δ,MFCC[0-14], MFCCΔ[0-5,7-12],Log Mel[2,3]*	Percentile	99.0	*F0Sub+Δ,ShimmerLocalΔ,JitterDDPΔ,PCMΔ,VoicingProb,MFCC[0,2,5-9,11, 12,14],MFCCΔ[1-8,10-14],LogMel[3,5],LogMelΔ[1,3-7], LSP[0,2-7], LSPΔ[1]*
		1.0	*F0 Env+Δ,MFCC[0-14], MFCCΔ[0-4,6-12],Log Mel[0,2,3,5-7]*		1.0	*F0EnvΔ,PCM+Δ,VoicingProb, LogMel[0,2,4,7], LogMelΔ [1,5-7], LSP[0,1,3-7], LSPΔ[0,1],MFCC[6,7,9,12,14],MFCCΔ[0-3,5,7,9,10,12-14]*
	Percentile range		*F0 Env+Δ,MFCC+Δ[0-14],Log Mel[0,2,3,5-7], LogMelΔ[2]*	Percentile range		*F0EnvΔ,VoicingProb+Δ, MFCC[0,4,8,9,11,12],MFCCΔ[0,3,5-8,10,12-14], LogMel[3-5,7],LogMelΔ[0,2-4], LSP [0-7], LSPΔ[1-6]*
	Up-level time		____	Up-level time	75	*F0SubΔ,F0Env, VoicingProb, Jitter LocalΔ, MFCC[0,3-6,8-11,13],MFCCΔ[0,2-11,14],LogMel[0-7],Log MelΔ[2,3,6, 7], LSP[0,1,2,4,5], LSPΔ[0,1,3,5-7]*
					90	*F0Sub,F0Env,ShimmerLocalΔ, Jitter DDP+Δ, VoicingProb, Jitter LocalΔ, MFCC[1,6,13,14], MFCCΔ[2], Log Mel[0,1,6,7], LSP[1,3-7], LSPΔ[5]*

References

[1] A. Kołakowska, A. Landowska, M. Szwoch, W. Szwoch, and M. R. Wróbel, "Emotion Recognition and Its Applications," *Advances in Intelligent Systems and Computing*. 2014.

[2] J. Nicholson, K. Takahashi, and R. Nakatsu, "Emotion recognition in speech using neural networks," *Neural Comput. Appl.*, 2000.

[3] R. Corive *et al.*, "Emotion recognition in human-computer interaction," *IEEE Signal Process. Mag.*, 2001.

[4] R. Banse and K. R. Scherer, "Acoustic Profiles in Vocal Emotion Expression," *J. Pers. Soc. Psychol.*, 1996.

[5] M. Schubiger, *English intonation, its form and function.* T�bingen: M. Niemeyer Verlag, 1958.

[6] J. D. O'Connor and G. F. Arnold, *Intonation of colloquial English*. London: Longman, 1973.

[7] D. Ververidis and C. Kotropoulos, "Emotional speech recognition: Resources, features, and methods," *Speech Commun.*, 2006.

[8] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognit.*, 2011.

[9] S. Ramakrishnan, "Recognition of Emotion from Speech: A Review," in *Speech Enhancement, Modeling and Recognition- Algorithms and Applications*, 2012.

[10] X. Cheng, X. Wang, T. Ouyang, and Z. Feng, "Advances in Emotion Recognition:

Link to Depressive Disorder," in *Neurological and Mental Disorders*, 2020.

[11] L. Meteyard, C. Fairfield, P. Sharp, H. Bettle, S. Philpott, and L. Wood, "Testing the ecological validity of an automated procedure for measuring speech intelligibility (icSpeech Intelligibility Scorer)." Internal report, Department of Clinical Language Sciences, University of Reading, 2014.

[12] G. C. Bunn, *The truth machine: A social history of the lie detector*. 2012.

[13] S. T. Saste and S. M. Jagdale, "Emotion recognition from speech using MFCC and DWT for security system," in *Proceedings of the International Conference on Electronics, Communication and Aerospace Technology, ICECA 2017*, 2017.

[14] A. Tawari and M. Trivedi, "Speech emotion analysis in noisy real-world environment," in *Proceedings - International Conference on Pattern Recognition*, 2010.

[15] S. Zepf, J. Hernandez, A. Schmitt, W. Minker, and R. W. Picard, "Driver Emotion Recognition for Intelligent Vehicles: A Survey," *ACM Computing Surveys*. 2020.

[16] C. M. Jones and I.-M. Jonsson, "Automatic recognition of affective cues in the speech of car drivers to allow appropriate responses," *North*, 2005.

[17] J. Luig and A. Sontacchi, "A speech database for stress monitoring in the cockpit," *Proc. Inst. Mech. Eng. Part G J. Aerosp. Eng.*, 2014.

[18] S. Lukose and S. S. Upadhya, "Music player based on emotion recognition of voice signals," in *2017 International Conference on Intelligent Computing, Instrumentation and Control Technologies, ICICICT 2017*, 2018.

[19] W. Li, Y. Zhang, and Y. Fu, "Speech emotion recognition in E-learning system based on affective computing," in *Proceedings - Third International Conference on Natural Computation, ICNC 2007*, 2007.

[20] J. G., D. Sundgren, R. Rahmani, A. Larsson, A. Moran, and I. Bonet, "Speech emotion recognition in emotional feedback for Human-Robot Interaction," *Int. J. Adv. Res. Artif. Intell.*, 2015.

[21] S. Ramakrishnan and I. M. M. El Emary, "Speech emotion recognition approaches in human computer interaction," *Telecommun. Syst.*, 2013.

[22] M. Anjum, "Emotion Recognition from Speech for an Interactive Robot Agent," in *Proceedings of the 2019 IEEE/SICE International Symposium on System Integration, SII 2019*, 2019.

[23] R. A. Koteswara, K. Swarna, and D. V. Hima, *Acoustic Modeling for Emotion Recognition*. 2016.

[24] N. Cummins, S. Scherer, J. Krajewski, S. Schnieder, J. Epps, and T. F. Quatieri, "A review of depression and suicide risk assessment using speech analysis," *Speech Communication*. 2015.

[25] T. M. Mitchell, *Machine Learning*, 1st ed. USA: McGraw-Hill, Inc., 1997.

[26] P. Domingos, "A few useful things to know about machine learning," *Communications of the ACM*. 2012.

[27] J. H. L. Hansen, "Analysis and compensation of speech under stress and noise for environmental robustness in speech recognition," *Speech Commun.*, 1996.

[28] M. Hagmüller, E. Rank, and G. Kubin, "Evaluation of the Human Voice for Indications of Workload-induced Stress in the Aviation Environment." 2006.

[29] S. E. Lively, D. B. Pisoni, W. Van Summers, and R. H. Bernacki, "Effects of cognitive workload on speech production: acoustic analyses and perceptual consequences," *J. Acoust. Soc. Am.*, vol. 93, no. 5, pp. 2962–2973, May 1993.

[30] D. J. Hermes, "Measurement of pitch by subharmonic summation," *J. Acoust. Soc. Am.*, 1988.

[31] B. Schuller, G. Rigoll, and M. Lang, "Hidden Markov model-based speech emotion recognition," in *Proceedings - IEEE International Conference on Multimedia and Expo*, 2003.

[32] B. Schuller, G. Rigol, and M. Lang, "Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine - Belief network architecture," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2004.

[33] K. P. Seng, L. M. Ang, and C. S. Ooi, "A Combined Rule-Based & Machine Learning Audio-Visual Emotion Recognition Approach," *IEEE Trans. Affect. Comput.*, 2018.

[34] M. Lugger and B. Yang, "The relevance of voice quality features in speaker independent emotion recognition," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2007.

[35] L. Xi *et al.*, "Stress and emotion classification using jitter and shimmer features," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing -*

Proceedings, 2007.

[36] K. I. Nordstrom, G. Tzanetakis, and P. F. Driessen, "Transforming perceived vocal effort and breathiness using adaptive pre-emphasis linear prediction," *IEEE Trans. Audio, Speech Lang. Process.*, 2008.

[37] V. Rozgić, S. Ananthakrishnan, S. Saleem, R. Kumar, A. N. Vembu, and R. Prasad, "Emotion recognition using acoustic and lexical features," in *13th Annual Conference of the International Speech Communication Association 2012, INTERSPEECH 2012*, 2012.

[38] M. J. Kim, J. Yoo, Y. Kim, and H. Kim, "Speech emotion classification using tree-structured sparse logistic regression," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2015.

[39] S. Ying and Z. Xue-Ying, "Characteristics of human auditory model based on compensation of glottal features in speech emotion recognition," *Futur. Gener. Comput. Syst.*, 2018.

[40] S. Vekkot and S. Tripathi, "Significance of glottal closure instants detection algorithms in vocal emotion conversion," in *Advances in Intelligent Systems and Computing*, 2018.

[41] T. L. Nwe, S. W. Foo, and L. C. De Silva, "Speech emotion recognition using hidden Markov models," *Speech Commun.*, 2003.

[42] D. Neiberg, K. Elenius, and K. Laskowski, "Emotion recognition in spontaneous speech using GMMs," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2006.

[43] H. K. Palo, M. N. Mohanty, and M. Chandra, "Use of different features for emotion recognition using MLP network," in *Advances in Intelligent Systems and Computing*, 2015.

[44] Y. CK *et al.*, "Bispectral features and mean shift clustering for stress and emotion recognition from natural speech," *Comput. Electr. Eng.*, 2017.

[45] Y. Wang and W. Hu, "Speech emotion recognition based on improved MFCC," in *ACM International Conference Proceeding Series*, 2018.

[46] H. M. Teager, "Some Observations on Oral Air Flow During Phonation," *IEEE Trans. Acoust.*, 1980.

[47] J. F. Kaiser, “Some useful properties of Teager’s energy operators,” in *Proceedings - ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing*, 1993.

[48] D. A. Cairns and H. L. John Hansen, “Nonlinear analysis and classification of speech under stressed conditions,” *J. Acoust. Soc. Am.*, 1994.

[49] G. Zhou, J. H. L. Hansen, and J. F. Kaiser, “Classification of speech under stress based on features derived from the nonlinear Teager energy operator,” in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 1998.

[50] G. Zhou, J. H. L. Hansen, and J. F. Kaiser, “Nonlinear feature based classification of speech under stress,” *IEEE Trans. Speech Audio Process.*, 2001.

[51] R. Sun and E. Moore, “Investigating glottal parameters and teager energy operators in emotion recognition,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2011.

[52] K. Wu, D. Zhang, and G. Lu, “GMAT: Glottal closure instants detection based on the Multiresolution Absolute Teager–Kaiser energy operator,” *Digit. Signal Process. A Rev. J.*, 2017.

[53] S. E. Bou-Ghazale and J. H. L. Hansen, “A Comparative Study of Traditional and Newly Proposed Features for Recognition of Speech under Stress,” *IEEE Trans. Speech Audio Process.*, 2000.

[54] A. Alpan, J. Schoentgen, Y. Maryn, and F. Grenez, “Automatic perceptual categorization of disordered connected speech,” in *Proceedings of the 11th Annual Conference of the International Speech Communication Association, INTERSPEECH 2010*, 2010.

[55] Y. Attabi, M. J. Alam, P. Dumouchel, P. Kenny, and D. O’Shaughnessy, “Multiple windowed spectral features for emotion recognition,” in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2013.

[56] J. Pohjalainen, R. Saeidi, T. Kinnunen, and P. Alku, “Extended weighted linear prediction (XLP) analysis of speech and its application to speaker verification in adverse conditions,” in *Proceedings of the 11th Annual Conference of the International Speech Communication Association, INTERSPEECH 2010*, 2010.

[57] Z. T. Liu, Q. Xie, M. Wu, W. H. Cao, Y. Mei, and J. W. Mao, "Speech emotion recognition based on an improved brain emotion learning model," *Neurocomputing*, 2018.

[58] S. Wu, T. H. Falk, and W.-Y. Chan, "Automatic speech emotion recognition using modulation spectral features," *Speech Commun.*, 2011.

[59] L. Zao, D. Cavalcante, and R. Coelho, "Time-frequency feature and AMS-GMM mask for acoustic emotion classification," *IEEE Signal Process. Lett.*, 2014.

[60] K. Wang, N. An, B. N. Li, Y. Zhang, and L. Li, "Speech emotion recognition using Fourier parameters," *IEEE Trans. Affect. Comput.*, 2015.

[61] S. Deb and S. Dandapat, "Multiscale Amplitude Feature and Significance of Enhanced Vocal Tract Information for Emotion Classification," *IEEE Trans. Cybern.*, 2019.

[62] L. Caponetti, C. A. Buscicchio, and G. Castellano, "Biologically inspired emotion recognition from speech," *EURASIP J. Adv. Signal Process.*, 2011.

[63] L. He, M. Lech, N. C. Maddage, and N. B. Allen, "Study of empirical mode decomposition and spectral analysis for stress and emotion classification in natural speech," *Biomed. Signal Process. Control*, 2011.

[64] T. Polzehl, A. Schmitt, F. Metze, and M. Wagner, "Anger recognition in speech using acoustic and linguistic cues," *Speech Commun.*, 2011.

[65] S. Ramamohan and S. Dandapat, "Sinusoidal model-based analysis and classification of stressed speech," *IEEE Trans. Audio, Speech Lang. Process.*, 2006.

[66] Y. C.K., M. Hariharan, R. Ngadiran, A. H. Adom, S. Yaacob, and K. Polat, "Hybrid BBO_PSO and higher order spectral features for emotion and stress recognition from natural speech," *Appl. Soft Comput. J.*, 2017.

[67] B. Yang and M. Lugger, "Emotion recognition from speech signals using new harmony features," *Signal Processing*, 2010.

[68] Y. Sun, G. Wen, and J. Wang, "Weighted spectral features based on local Hu moments for speech emotion recognition," *Biomed. Signal Process. Control*, 2015.

[69] S. Motamed, S. Setayeshi, and A. Rabiee, "Speech emotion recognition based on a modified brain emotional learning model," *Biol. Inspired Cogn. Archit.*, 2017.

[70] H. M. Teager and S. M. Teager, "Evidence for Nonlinear Sound Production Mechanisms in the Vocal Tract," in *Speech Production and Speech Modelling*, 1990.

[71] J. F. Kaiser, "On a simple algorithm to calculate the 'energy' of a signal," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 1990.

[72] L. S. A. Low, N. C. Maddage, M. Lech, L. Sheeber, and N. Allen, "Influence of acoustic low-level descriptors in the detection of clinical depression in adolescents," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2010.

[73] L. S. A. Low, N. C. Maddage, M. Lech, L. B. Sheeber, and N. B. Allen, "Detection of clinical depression in adolescents' speech during family interactions," *IEEE Trans. Biomed. Eng.*, 2011.

[74] K. E. B. Ooi, M. Lech, and N. B. Allen, "Multichannel weighted speech classification system for prediction of major depression in adolescents," *IEEE Trans. Biomed. Eng.*, 2013.

[75] K. E. B. Ooi, M. Lech, and N. Brian Allen, "Prediction of major depression in adolescents using an optimized multi-channel weighted speech classification system," *Biomed. Signal Process. Control*, 2014.

[76] B. Schuller, S. Steidl, and A. Batliner, "The INTERSPEECH 2009 emotion challenge," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2009.

[77] B. Schuller *et al.*, "The INTERSPEECH 2010 paralinguistic challenge," in *Proceedings of the 11th Annual Conference of the International Speech Communication Association, INTERSPEECH 2010*, 2010.

[78] W. A. Jassim, R. Paramesran, and N. Harte, "Speech emotion classification using combined neurogram and INTERSPEECH 2010 paralinguistic challenge features," *IET Signal Process.*, 2017.

[79] P. W. Hsiao and C. P. Chen, "Effective Attention Mechanism in Dynamic Models for Speech Emotion Recognition," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2018.

[80] R. D. Fonnegra and G. M. Díaz, "Speech emotion recognition integrating paralinguistic features and auto-encoders in a deep learning model," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes*

in Bioinformatics), 2018.

[81] A. Rayaluru, S. R. Bandela, and T. Kishore Kumar, "Speech emotion recognition using feature selection with adaptive structure learning," in *Proceedings - 2019 IEEE International Symposium on Smart Electronic Systems, iSES 2019*, 2019.

[82] M. E. Wall, A. Rechtsteiner, and L. M. Rocha, "Singular Value Decomposition and Principal Component Analysis," in *A Practical Approach to Microarray Data Analysis*, 2005.

[83] C. Bartenhagen, H. U. Klein, C. Ruckert, X. Jiang, and M. Dugas, "Comparative study of unsupervised dimension reduction techniques for the visualization of microarray gene expression data," *BMC Bioinformatics*, 2010.

[84] C. M. Lee and S. S. Narayanan, "Toward detecting emotions in spoken dialogs," *IEEE Trans. Speech Audio Process.*, 2005.

[85] D. Wu, T. D. Parsons, and S. S. Narayanan, "Acoustic feature analysis in speech emotion primitives estimation," in *Proceedings of the 11th Annual Conference of the International Speech Communication Association, INTERSPEECH 2010*, 2010.

[86] S. Zhang, L. Li, and Z. Zhao, "Spoken emotion recognition using kernel discriminant locally linear embedding," *Electron. Lett.*, 2010.

[87] P. Song, S. Ou, W. Zheng, Y. Jin, and L. Zhao, "Speech emotion recognition using transfer non-negative matrix factorization," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2016.

[88] S. Sahu, R. Gupta, G. Sivaraman, W. AbdAlmageed, and C. Espy-Wilson, "Adversarial auto-encoders for speech based emotion recognition," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2017.

[89] S. Latif, R. Rana, J. Qadir, and J. Epps, "Variational autoencoders to learn latent representations of speech emotion," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2018.

[90] S. H. Chen, J. C. Wang, W. C. Hsieh, Y. H. Chin, C. W. Ho, and C. H. Wu, "Speech emotion classification using multiple kernel Gaussian process," in *2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA 2016*, 2017.

[91] S. Zhang and X. Zhao, “Dimensionality reduction-based spoken emotion recognition,” *Multimed. Tools Appl.*, 2013.

[92] J. C. Ang, A. Mirzal, H. Haron, and H. N. A. Hamed, “Supervised, unsupervised, and semi-supervised feature selection: A review on gene selection,” *IEEE/ACM Trans. Comput. Biol. Bioinforma.*, 2016.

[93] A. Arruti, I. Cearreta, A. Álvarez, E. Lazkano, and B. Sierra, “Feature selection for speech emotion recognition in Spanish and Basque: On the use of machine learning to improve human-computer interaction,” *PLoS One*, 2014.

[94] T. Özseven, “A novel feature selection method for speech emotion recognition,” *Appl. Acoust.*, 2019.

[95] L. Sun, S. Fu, and F. Wang, “Decision tree SVM model with Fisher feature selection for speech emotion recognition,” *Eurasip J. Audio, Speech, Music Process.*, 2019.

[96] S. Kuchibhotla, H. D. Vankayalapati, and K. R. Anne, “An optimal two stage feature selection for speech emotion recognition using acoustic features,” *Int. J. Speech Technol.*, 2016.

[97] Y. Jin, P. Song, W. Zheng, and L. Zhao, “A feature selection and feature fusion combination method for speaker-independent speech emotion recognition,” in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2014.

[98] J. Yan, X. Wang, W. Gu, and L. Ma, “Speech emotion recognition based on sparse representation,” *Arch. Acoust.*, 2013.

[99] S. Zhang, X. Zhao, and B. Lei, “Speech emotion recognition using an enhanced kernel isomap for human-robot interaction,” *Int. J. Adv. Robot. Syst.*, 2013.

[100] A. P. Gudmalwar, C. V. Rama Rao, and A. Dutta, “Improving the performance of the speaker emotion recognition based on low dimension prosody features vector,” *Int. J. Speech Technol.*, 2019.

[101] Z. W. Huang, W. T. Xue, and Q. R. Mao, “Speech emotion recognition with unsupervised feature learning,” *Front. Inf. Technol. Electron. Eng.*, 2015.

[102] C. M. Bishop, *Machine Learning and Pattern Recogniton*. 2006.

[103] H. Hu, M. X. Xu, and W. Wu, “GMM supervector based SVM with spectral features for speech emotion recognition,” in *ICASSP, IEEE International Conference on*

Acoustics, Speech and Signal Processing - Proceedings, 2007.

[104] A. Georgogiannis and V. Digalakis, "Speech Emotion Recognition using non-linear Teager energy based features in noisy environments," in *European Signal Processing Conference*, 2012.

[105] X. Zhao, S. Zhang, and B. Lei, "Robust emotion recognition in noisy speech via sparse representation," *Neural Comput. Appl.*, 2014.

[106] M. Bashirpour and M. Geravanchizadeh, "Speech emotion recognition based on power normalized cepstral coefficients in noisy conditions," *Iran. J. Electr. Electron. Eng.*, 2016.

[107] J. Xiaoqing, X. Kewen, L. Yongliang, and B. Jianchuan, "Noisy speech emotion recognition using sample reconstruction and multiple-kernel learning," *J. China Univ. Posts Telecommun.*, 2017.

[108] Y. Huang, W. Ao, and G. Zhang, "Novel Sub-band Spectral Centroid Weighted Wavelet Packet Features with Importance-Weighted Support Vector Machines for Robust Speech Emotion Recognition," *Wirel. Pers. Commun.*, 2017.

[109] S. Sekkate, M. Khalil, A. Adib, and S. Ben Jebara, "An investigation of a feature-level fusion for noisy speech emotion recognition," *Computers*, 2019.

[110] D. Pearce and H. G. Hirsch, "The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *6th International Conference on Spoken Language Processing, ICSLP 2000*, 2000.

[111] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A database of German emotional speech," in *9th European Conference on Speech Communication and Technology*, 2005.

[112] C. Busso *et al.*, "IEMOCAP: Interactive emotional dyadic motion capture database," *Lang. Resour. Eval.*, 2008.

[113] R. O. Duda, P. E. Hart, and D. G. Stork, "Pattern classification," *New York John Wiley, Sect.*, 2001.

[114] G. Paraskevopoulos, E. Tzinis, N. Ellinas, T. Giannakopoulos, and A. Potamianos, "Unsupervised low-rank representations for speech emotion recognition," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2019.

[115] R. Aziz, C. K. Verma, and N. Srivastava, "Dimension reduction methods for microarray data: a review," *AIMS Bioeng.*, 2017.

[116] T. Ozseven, "Evaluation of the Effect of Frame Size on Speech Emotion Recognition," in *ISMSIT 2018 - 2nd International Symposium on Multidisciplinary Studies and Innovative Technologies, Proceedings*, 2018.

[117] S. Chakroborty, A. Roy, and G. Saha, "Fusion of a Complementary Feature Set with MFCC for Improved Closed Set Text-Independent Speaker Identification," in *Proceedings of the IEEE International Conference on Industrial Technology*, 2006.

[118] T. Quatieri -, *Discrete-Time Speech Signal Processing: Principles and Practice*, First. USA: Prentice Hall Press, 2001.

[119] K. S. Rao, V. R. Reddy, and S. Maity, *Language Identification Using Spectral and Prosodic Features*, 1st ed. Springer International Publishing, 2015.

[120] S. Alghowinem *et al.*, "A comparative study of different classifiers for detecting depression from spontaneous speech," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2013.

[121] H. Jiang *et al.*, "Investigation of different speech types and emotions for detecting depression using different classifiers," *Speech Commun.*, 2017.

[122] C. Ding, T. Li, and M. I. Jordan, "Convex and semi-nonnegative matrix factorizations," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2010.

[123] N. Gillis and A. Kumar, "Exact and heuristic algorithms for semi-nonnegative matrix factorization," *SIAM J. Matrix Anal. Appl.*, 2015.

[124] F. Eyben, M. Wöllmer, and B. Schuller, "OpenSMILE - The Munich versatile and fast open-source audio feature extractor," in *MM'10 - Proceedings of the ACM Multimedia 2010 International Conference*, 2010.

[125] X. Valero and F. Alias, "Gammatone cepstral coefficients: Biologically inspired features for non-speech audio classification," *IEEE Trans. Multimed.*, 2012.

[126] J. Guo, Y. Quo, X. Kong, and R. He, "Unsupervised feature selection with ordinal locality," in *Proceedings - IEEE International Conference on Multimedia and Expo*, 2017.

[127] L. Du and Y. D. Shen, "Unsupervised feature selection with adaptive structure learning," in *Proceedings of the ACM SIGKDD International Conference on*

Knowledge Discovery and Data Mining, 2015.

[128] W. Jiang, Z. Wang, J. S. Jin, X. Han, and C. Li, "Speech emotion recognition with heterogeneous feature unification of deep neural network," *Sensors (Switzerland)*, 2019.

[129] C. Kim and R. M. Stern, "Power-Normalized Cepstral Coefficients (PNCC) for Robust Speech Recognition," *IEEE/ACM Trans. Audio Speech Lang. Process.*, 2016.

[130] S. Surendran and T. K. Kumar, "Oblique projection and cepstral subtraction in signal subspace speech enhancement for colored noise reduction," *IEEE/ACM Trans. Audio Speech Lang. Process.*, 2018.

[131] M. Sun, Y. Li, J. F. Gemmeke, and X. Zhang, "Speech enhancement under low SNR conditions via noise estimation using sparse and low-rank NMF with Kullback-Leibler divergence," *IEEE Trans. Audio, Speech Lang. Process.*, 2015.

[132] S. U. N. Wood and J. Rouat, "Unsupervised Low Latency Speech Enhancement With RT-GCC-NMF," *IEEE J. Sel. Top. Signal Process.*, 2019.

[133] N. Lyubimov and M. Kotov, "Non-negative matrix factorization with linear constraints for single-channel speech enhancement," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2013.

[134] B. Gold, N. Morgan, and D. Ellis, *Speech and Audio Signal Processing: Processing and Perception of Speech and Music: Second Edition*. 2011.

[135] R. J. Mcauly and T. F. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Trans. Acoust.*, 1986.

List of Publications

International Journals:

[1] Surekha Reddy Bandela and T. Kishore Kumar, "**Unsupervised Feature Selection and NMF Denoising for Robust Speech Emotion Recognition**", **Applied Acoustics Journal**. DoI:10.1016/j.apacoust.2020.107645, 172 (2021) 107645, PP. 1-15, September 2020. **(SCI)**

[2] Surekha Reddy Bandela and T. Kishore Kumar, "**Speech Emotion Recognition using Unsupervised Feature Selection Algorithms**", **Radioengineering Journal**, DoI: 10.13164/re.2020.0353, Vol. 29, No. 2, PP. 353-364, June 2020. **(SCI)**

[3] Surekha Reddy Bandela and T. Kishore Kumar, **"Speech emotion recognition using semi-NMF feature optimization", Turkish Journal of Electrical Engineering & Computer Sciences,** DoI: 10.3906/elk-1903-121, Vol 27, PP. 3741-3757, June 2019.**(SCI)**

International Conferences:

[1] Surekha Reddy Bandela and T. Kishore Kumar, " **Emotion Recognition of Stressed Speech using Teager Energy and Linear Prediction Features**", 18th International Conference on Advanced Learning Technologies (ICALT), IIT Bombay, July 9-13, 2018. (**IEEE**)

[2] Surekha Reddy Bandela and Kishore Kumar T., "**Stressed Speech Emotion Recognition using feature fusion of Teager Energy Operator and MFCC**", 8th International Conference On Computing, Communication and Networking Technologies (ICCCNT-2017), IEEE, July 3-5, IIT Delhi. (**IEEE**)