

SPEECH BASED EMOTION RECOGNITION USING NON-STATIONARY DECOMPOSITIONS AND OPTIMAL FEATURES

**A Thesis Submitted
In Partial Fulfillment of the Requirements
for the Degree of**

DOCTOR OF PHILOSOPHY

by

Ravi

(Roll No. 2K20/PHDEC/12)

Under the Supervision of

Dr. Sachin Taran

**(Assistant Professor, Department of Electronics and Communication Engineering
/ Delhi Technological University)**



Department of Electronics and Communication Engineering

**DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road, Delhi 110042, India**

April, 2025

ACKNOWLEDGEMENT

I would like to extend my gratitude and my sincere thanks to my honorable, esteemed supervisor, **Dr. Sachin Taran**. He is not only a great professor with deep vision but also, most importantly, a kind person. I sincerely thank him for his exemplary guidance and encouragement. His trust and support inspired me in the moments when I needed to make the right decisions, and I am glad to work with him.

I am also thankful to the entire faculty and staff of the electronics and communication engineering department of Delhi Technological University, New Delhi for their unyielding encouragement.

I am greatly indebted to all my friends, who have graciously applied themselves to the task of helping me with ample moral support and valuable suggestions. A well-deserved expression of appreciation goes to my father, Mr. A.N Pandey, my mother Ms. Lalita Pandey, my wife, Ms. Shraddha, and my son, Avyan Pandey, for their support and prayers.

Ravi

CANDIDATE'S DECLARATION

I **RAVI** hereby certify that the work which is being presented in the thesis entitled **Speech Based Emotion Recognition Using Non-Stationary Decompositions and Optimal Features** in partial fulfillment of the requirements for the award of the Degree of Doctor of Philosophy, submitted in the department of **Electronics and Communication Engineering, Delhi Technological University** is an authenticated record of my own work carried out during the period from **August 2020 to December 2024** under the supervision of **Dr. Sachin Taran**.

The matter discussed in the thesis has not been submitted by me for the award of any other degree of this or any other institute.

Candidate's Signature

This is to certify that the student has incorporated all the corrections suggested by the examiner in the thesis and the statement made by the candidate is correct to the best of our knowledge.

Signature of Supervisor

Signature of External Examiner

CERTIFICATE BY THE SUPERVISOR

Certified that **Ravi (2k20/PHDEC/12)** has carried out their research work presented in this thesis entitled “**SPEECH BASED EMOTION RECOGNITION USING NON-STATIONARY DECOMPOSITIONS AND OPTIMAL FEATURES**” for the award of **Doctor of Philosophy** from Department of Electronics and Communication Engineering, Delhi Technological University, Delhi, under my supervision. The thesis embodies results of original work, and studies are carried out by the student himself and the contents of the thesis do not form the basis for the award of any other degree to the candidate or to anybody else from this or any other University/Institution.

Dr. Sachin Taran
Supervisor
Department of ECE
Delhi Technological University
Delhi-110042, India

Place:

Date:

Speech Based Emotion Recognition Using Non-Stationary Decompositions and Optimal Features

Ravi

ABSTRACT

Speech emotion recognition (SER) is the process of identifying and classifying emotions expressed in spoken language using audio features and computational models. It has applications in e-learning, robotic interfaces, computer games, entertainment, audio surveillance, clinical studies, and more. Despite its promising applications, emotion recognition from speech signals is a challenging domain due to the inherent non-stationary and multicomponent nature of speech and language sensitive SERs. The complete SER framework is broadly divided into three stages: (i) signal preprocessing, (ii) feature extraction and selection, and (iii) classification. Contributions at any stage can improve the performance of SER models. This research work focuses on proposing a non-stationary signal processing framework for speech signals and evaluates the performance of the SER models for binary class, multiclass and multilingual scenarios.

Speech signals are non-stationary, meaning their statistical properties change over time. In this thesis, initially a rational dilation wavelet transform based method is presented to analyze speech signal for the binary SER system. The wavelet improves signal predictability and provides better time-frequency analysis. In this work four basic features complexity, average amplitude change, mobility and zero crossing rate are extracted. The proposed framework tested on publically available RAVDESS dataset and achieves 83.30% accuracy for binary class emotion. Even though the wavelet-based method offers better time-frequency representation, it is observed that wavelet transforms can be less effective for analyzing nonlinear signals due to their reliance on predefined basis functions. To overcome this issue, Empirical Mode Decomposition (EMD) is explored. The data-driven nature of EMD allows it to provide a more intuitive and interpretable representation of the signal's components, capturing subtle nuances and variations. Additionally, EMD's ability to perform localized time-frequency analysis ensures precise identification of transient features and other detailed structures within the speech signal. In this framework ratio feature based on energy and statistical measures are calculated from MFCC coefficients. The proposed EMD based SER framework is tested on EMOVO and RAVDESS dataset and achieves 95.30 and 90.01% accuracy respectively for binary speech emotion classification.

EMD provides a better SER performance compared to wavelets. However, the recursive nature of EMD creates mode mixing and it is also noise sensitive. To address this issue, a Variational Mode Decomposition (VMD) based method is adopted. One

significant benefit of VMD is its enhanced noise resilience, which minimizes mode mixing and leads to more stable decompositions of the signal. It also provides flexibility to control decomposition parameters. VMD with Teager-Kaiser energy operator-based features are explored for binary class and multiclass emotion classification. Meanwhile, the statistical measures (such as mean and variance) of MFCC and pitch frequency are computed after framing the preprocessed signal to account for temporal variations in speech. The introduced method is tested on RAVDESS dataset and provides better performance for binary class emotion recognition as compared to existing works. Although, the accuracy of the model reduces as the number of emotion classes increased.

To address multiclass emotion recognition issue energy based dominant VMD modes selection framework is proposed. First, VMD separates the speech signal into multiple modes where each represents distinct frequency components. The energy of each mode is then calculated to identify the dominant modes that contribute most significantly to the speech signal's characteristics. These dominant modes are subsequently used for signal reconstruction. Finally, the reconstructed signal is used for the feature extraction and emotion classification. In this framework spectral and prosodic features like Mel spectrum, spectral crest, spectral entropy, spectral kurtosis, spectral centroid, mel-frequency cepstral coefficients and their derivatives, gammatone cepstral coefficients and Pitch frequency. The proposed SER framework is tested on RAVDESS, EMOVO, Emo-DB and IEMOCAP dataset and achieves 93.80%, 93.40, 95.08% and 83.10% accuracy respectively. However, noise in raw speech signals could still obscure subtle emotional features, which can degrade performance. Additionally, its dependence on predefined parameters for mode tuning might not generalize well to highly variable or noisy inputs, especially for multilingual speech emotion recognition (MLSER).

To address previous work limitations, an enhanced signal preprocessing framework for MLSER is proposed. In this framework, silence is removed using short-time energy and spectral centroid, ensuring that only relevant speech segments are processed. VMD is then applied for signal decomposition, with an improved Bhattacharyya distance guiding mode tuning for noise removal. Finally, the denoised signal is used for feature extraction and emotion classification. In this framework, spectral and prosodic features such as the Mel spectrum, spectral crest, spectral entropy, spectral kurtosis, spectral centroid, Mel-frequency cepstral coefficients and their derivatives, gammatone cepstral coefficients, and pitch frequency are used. This framework is tested on the RAVDESS, EMOVO, and Emo-DB datasets, and a multilingual dataset is created by combining these three datasets. The proposed MLSER model achieved 93.4% accuracy for multilingual and multiclass dataset.

This thesis presents a progression of methods for SER with addressing the challenges of non-stationary and noisy speech signals. Initially, a wavelet-based approach

using rational dilation wavelet transform achieved 83.30% accuracy for binary emotions but struggled with highly nonlinear signals. EMD improved results to 90.01% accuracy yet suffered from noise sensitivity and mode mixing. VMD overcame these limitations, enhancing noise resilience and achieving 100% accuracy for binary class emotion recognition. Further VMD explored for multiclass SER, employing energy-based dominant mode selection to achieve 95.08% accuracy. VMD based speech pre-processing is improved by silence removal and refined noise handling. The proposed approach is tested on multilingual SER framework and achieved 93.4% accuracy. This thesis presents the SER framework for binary class, multiclass, and multilingual emotion identification, and proposed works outperform other existing state-of-the-art frameworks.

LIST OF PUBLICATION

List of Papers in Journals

- **Published**

- [1] Ravi and S. Taran, "A nonlinear feature extraction approach for speech emotion recognition using VMD and TKEO." *Applied Acoustics*, Elsevier, vol. 214, pp. 109667, 2023.
- [2] Ravi and S. Taran , "A novel decomposition-based architecture for multilingual speech emotion recognition." *Neural Computing and Applications*, Springer, vol. 36, pp. 9347-9359, 2024.
- [3] Ravi and S. Taran, "Emotion Recognition Using Energy Based Adaptive Mode Selection." *Speech Communication*, Elsevier, pp.103228, 2025.

- **Communicated**

- [1] Ravi and S. Taran, " A Filtering Approach for Speech Emotion Recognition Using Wavelet Approximation Coefficient." *Measurement*, Elsevier. (Preparing 2nd revision)
- [2] Ravi and S. Taran, " Multi Level Filltering Approach For Speech Emotion Recognition Using Vmd-Bi-Lstm." *Circuits, Systems & Signal processing*, Springer. (Submitted)

List of Papers in Conferences

- [1] Ravi and S. Taran, "Emotion Recognition Using Rational Dilation Wavelet Transform for Speech Signal," 2021 7th International Conference on Signal Processing and Communication (ICSC), IEEE, Noida, India, 2021, pp. 156-160.
- [2] Ravi and S. Taran, "Emotion Recognition in Speech Using MFCC and Energy Based Ratio Features," 2024 11th International Conference on Signal Processing and Integrated Networks (SPIN), IEEE, Noida, India, 2024, pp. 367-371.

List of Book Chapters

- [1] Hosain, Mehrab, Anukul Pandey, Sachin Taran, Ravi, Asmar Hafeez, and Nikhil. "Speech emotion recognition using empirical wavelet transform and cubic support vector machine." In *Artificial Intelligence: A tool for effective diagnostics*, pp. 13-1. Bristol, UK: IOP Publishing, 2024.

TABLE OF CONTENTS

| | |
|--|-----|
| LIST OF TABLES | xi |
| LIST OF FIGURES | xii |
| LIST OF ABBREVIATIONS | xiv |
| CHAPTER 1 INTRODUCTION AND LITERATURE SURVEY | 1 |
| 1.1 Background | 1 |
| 1.2 Speech Emotion Recognition | 2 |
| 1.3 Motivation | 3 |
| 1.4 Literature Survey | 4 |
| 1.4.1 Speech Preprocessing | 4 |
| 1.4.2 Feature Extraction | 6 |
| 1.4.3 Feature Selection | 9 |
| 1.4.4 Performance Matrices | 12 |
| 1.4.5 Overview of SER Techniques | 13 |
| 1.5 Research Gap and Problem Identification | 17 |
| 1.6 Research Objectives | 17 |
| 1.7 SER Database | 18 |
| 1.7.1 RAVDESS Emotion Database | 19 |
| 1.7.2 Emo-DB Database | 20 |
| 1.7.3 IEMOCAP Database | 20 |
| 1.7.4 EMOVO Database | 21 |
| 1.7.5 Multilingual Database | 22 |
| 1.8 Contribution | 22 |
| 1.8.1 Binary Class Speech Emotion Recognition | 23 |
| 1.8.2 Speech Emotion Recognition Using VMD-TKEO | 23 |
| 1.8.3 Multiclass Speech Emotion Recognition | 23 |
| 1.8.4 Multilingual Speech Emotion Recognition | 24 |
| 1.9 Organization of the Thesis | 24 |
| CHAPTER 2 BINARY CLASS SPEECH EMOTION RECOGNITION | 26 |
| 2.1 RDWT Based SER Model | 26 |
| 2.1.1 Dataset | 28 |
| 2.1.2 Preprocessing | 28 |
| 2.1.2.1 Rational Dilation Wavelet Transforms | 29 |
| 2.1.3 Feature Extraction | 30 |
| 2.1.4 Classification Models | 30 |
| 2.1.5 Results and Discussion | 30 |
| 2.2 SER Model based on Empirical Mode Transform | 34 |
| 2.2.1 Dataset | 34 |
| 2.2.2 EMD Decomposition | 35 |

| | |
|--|-----|
| 2.2.3 Feature Extraction | 37 |
| 2.2.3.1 Ratio Feature | 37 |
| 2.2.3.2 MFCC Feature | 38 |
| 2.2.4 Feature Selection and Classification | 39 |
| 2.2.5 Results and Discussion | 39 |
| 2.3 Comparison of RDWT and EMD based SER Model | 42 |
| 2.4 Summary | 42 |
| CHAPTER 3 SPEECH EMOTION RECOGNITION USING VARIATIONAL MODE DECOMPOSITION | 44 |
| 3.1 SER Model Using VMD-TKEO | 44 |
| 3.1.1 Variational Mode Decomposition | 45 |
| 3.1.2 Teager-Kaiser Energy Operator | 47 |
| 3.1.3 Feature Extraction | 48 |
| 3.1.3.1 Energy | 48 |
| 3.1.3.2 Pitch Frequency | 49 |
| 3.1.3.3 Mel Frequency Cepstrum Coefficients | 49 |
| 3.1.3.4 Statistical Measure | 50 |
| 3.1.4 Feature Selection | 51 |
| 3.1.5 Classification | 51 |
| 3.2 Results and Discussion | 51 |
| 3.3 Summary | 57 |
| CHAPTER 4 MULTICLASS SPEECH EMOTION RECOGNITION | 58 |
| 4.1 Multiclass SER Model | 59 |
| 4.1.1 Signal Preprocessing | 60 |
| 4.1.2 Feature Extraction, Selection and Classification | 63 |
| 4.2 Results and Discussion | 65 |
| 4.3 Summary | 73 |
| CHAPTER 5 MULTILINGUAL SPEECH EMOTION RECOGNITION | 74 |
| 5.1 Multilingual SER Model | 75 |
| 5.1.1 Silence Removal | 75 |
| 5.1.2 Noise Removal | 77 |
| 5.1.3 Feature Extraction & Classification | 81 |
| 5.2 Results and Discussion | 82 |
| 5.3 Summary | 90 |
| CHAPTER 6 CONCLUSION, FUTURE SCOPE AND SOCIAL IMPACT | 91 |
| 6.1 Conclusion | 91 |
| 6.2 Future Scope | 92 |
| 6.3 Social Impact | 93 |
| REFERENCES | 94 |
| LIST OF PUBLICATION AND THEIR PROOFS | 103 |
| PLAGIARISM REPORT | 112 |
| CURRICULUM VITAE | 114 |

LIST OF TABLES

| | |
|--|----|
| Table 2.1 KW Test P-Values of SB-Wise Extracted Features. | 31 |
| Table 2.2 Proposed Methods Classification Accuracies with DT, KNN and Ensemble Classifiers Variants. | 31 |
| Table 2.3 Performance Comparison with Existing State of Art. | 32 |
| Table 2.4 Precision, Recall and F1 Score of Proposed Model for EMOVO dataset. | 40 |
| Table 3.1 Total Features Extracted from Each Speech Signal. | 52 |
| Table 3.2 Binary Class Emotions Accuracy Comparison for VMD and EMD. | 52 |
| Table 3.3 Proposed Model Accuracy for Three Class Emotions. | 54 |
| Table 3.4 Proposed Model Accuracy for Four Class Emotions. | 54 |
| Table 3.5 Emotion Recognition Accuracy for the Different Test Sets. | 55 |
| Table 3.6 Emotion Recognition Results from Raw Speech Signal and VMD-TKEO Preprocessed Speech Signal. | 55 |
| Table 3.7 Comparison Table of Proposed Method with Existing Methods. | 56 |
| Table 4.1 For Example, Energy of Angry and Sad Emotion from four Datasets RAVDESS, EMOVO, Emo-DB, and IEMOCAP. E1 to E9 Represents the Energy of Modes from One to Nine. | 62 |
| Table 4.2 Precision, Recall and F1 Score for the RAVDESS Experiment. | 69 |
| Table 4.3 Precision, Recall and F1 Score for the EMOVO Experiment. | 69 |
| Table 4.4 Precision, Recall and F1 Score for the EMO-DB Experiment. | 69 |
| Table 4.5 Precision, Recall and F1 Score for the IEMOCAP Experiment. | 70 |
| Table 4.6 Recognition Rates of Various SER State of the Art Methods for Comparative Performance Analysis. | 70 |
| Table 4.7 Recognition Rates of Various SER State of the Art Methods for Elicited Dataset IEMOCAP for Comparative Performance Analysis. | 72 |
| Table 5.1 Improved Bhattacharya Distance for 9 Modes and 15 Modes Decomposition of Speech Signal. | 78 |
| Table 5.2 Maximum Distance of Adjacent Mode. | 80 |
| Table 5.3 Experimental Result of Proposed Method for RAVDESS Database. | 86 |
| Table 5.4 Experimental Result of Proposed Method for EMO-DB Database. | 86 |
| Table 5.5 Experimental Result of Proposed Method for EMOVO Database. | 87 |
| Table 5.6 Experimental Result of Proposed Method for Multilingual Database. | 87 |
| Table 5.7 Performance Comparison Benchmark for Individual Database. | 89 |
| Table 5.8 Performance Comparison Benchmark for Multilingual Database. | 90 |

LIST OF FIGURES

| | |
|---|----|
| Fig. 1.1 Speech Preprocessing Methods. | 4 |
| Fig. 1.2 Speech Feature Categories. | 7 |
| Fig. 1.3 Feature Selection Methods. | 10 |
| Fig. 1.4 Emotion Distribution in Percentage for RAVDESS Database. | 19 |
| Fig. 1.5 Emotion Distribution in Percentage for Emo-DB Database. | 20 |
| Fig. 1.6 Emotion Distribution in Percentage for IEMOCAP Database. | 21 |
| Fig. 1.7 Emotion Distribution in Percentage for EMOVO Database. | 21 |
| Fig. 1.8 Emotion Distribution in Percentage for Multilingual Database. | 22 |
| Fig. 2.1 Classification of Emotions Using RDWT Based Features. | 27 |
| Fig. 2.2 Example of Raw Speech Signal. | 27 |
| Fig. 2.3 Example of Preprocessed Speech Signal. | 28 |
| Fig. 2.4 Confusion Matrix of OE Classifier. | 32 |
| Fig. 2.5 ROC Curve of OE Classifier. | 33 |
| Fig. 2.6 Process Flow of Proposed SER Model. | 34 |
| Fig. 2.7 Speech Signal and IMFs after EMD Decomposition. | 36 |
| Fig. 2.8 Schema of MFCC Feature Extraction. | 38 |
| Fig. 2.9 Relieff Algorithm Predictor Rank Bar for Feature Selection. | 39 |
| Fig. 2.10 Region of Convergence of Proposed Framework. | 40 |
| Fig. 3.1 Overall Block Diagram of Proposed SER Architecture Based on VMD. | 45 |
| Fig. 3.2 Overall Block Diagram of Proposed SER Architecture Based on EMD. | 45 |
| Fig. 3.3 Spectrogram of Raw Speech Signal in Angry Emotion and VMD-TKEO Preprocessed Signal. | 47 |
| Fig. 3.4 Schema of MFCC Features Extraction for VMD-TKEO Method. | 49 |
| Fig. 4.1 Block Diagram of Proposed Framework. Where, K is Number of Modes. | 59 |
| Fig. 4.2 The Spectral Entropy Comparison of the Raw Speech Signal and the Reconstructed Signal. | 61 |
| Fig. 4.3 Predictor Rank Bar of RAVDESS Experiment. | 65 |
| Fig. 4.4 Recognition Rate of RAVDESS, Emo-DB and EMOVO Experiment. | 66 |
| Fig. 4.5 Confusion Matrix of RAVDESS Experiment. | 67 |
| Fig. 4.6 Confusion Matrix of EMOVO Experiment. | 67 |
| Fig. 4.7 Confusion Matrix of Emo-DB Experiment. | 68 |
| Fig. 4.8 Confusion Matrix of IEMOCAP Experiment. | 68 |
| Fig. 5.1 (a) Block Diagram of Signal Reconstruction. | 75 |
| Fig. 5.2 Raw Speech Signal and Speech Signal after Silence Removal. | 76 |
| Fig. 5.3 IBD for 9 Modes to 15 Modes Decomposition. | 78 |
| Fig. 5.4 Mean and Standard Deviation (SD) Plot for Thresholding. | 79 |
| Fig. 5.5 Speech Signal after Silence Removal and Noise Removal. | 79 |
| Fig. 5.6 Predictor Importance Weight Bar. | 82 |
| Fig. 5.7 Recognition Rate of Proposed Method for All Four Databases. | 83 |

| | |
|--|----|
| Fig. 5.8 Confusion Matrix for Italian Language Based EMOVO Database. | 83 |
| Fig. 5.9 Confusion Matrix for Emo-DB Database. | 84 |
| Fig. 5.10 Confusion Matrix for RAVDESS Database. | 84 |
| Fig. 5.11 Confusion Matrix for Multilingual Database. | 85 |
| Fig. 5.12 The Class-Level Balance Accuracy for RAVDESS, Emo-DB, EMOVO and Multilingual Database. | 85 |

LIST OF ABBREVIATIONS

AM-FM - Amplitude and Frequency Modulated
AI - Artificial Intelligence
AE - Autoencoder
AAC - Average Amplitude Change
BFN - Backpropagation-Free Network
Emo-DB - Berlin Emotional Database
BD - Bhattacharyya Distance
CAN - Categorical Neural Architecture
CNN - Convolution Neural Network
TEO-CB-Auto-Env - Critical Band Based TEO Autocorrelation Envelope Area
DFT - Discrete Fourier Transform
EMD - Empirical Mode Decomposition
GTCC - Gammatone Cepstral Coefficients
HW - Hamming Window
HBN - Hierarchical Bayesian Network
HT - Hilbert Transform
HCI - Human Computer Interaction
IMF - Intrinsic Mode Functions
INCA - Iterated Neighborhood Component Analysis
KNN - K-Nearest Neighbor
KW - Kruskal-Wallis
LDA - Linear Discriminant Analysis
LPC - Linear Prediction Coefficients
LSTM - Long Short-Term Memory
MFCC - Mel Frequency Cepstral Coefficients
MFMC - Mel Frequency Magnitude Coefficient
M-MFCC - Modified Mel Frequency Cepstral Coefficients
MLSER - Multilingual Speech Emotion Recognition
RDWT - Rational Dilation Wavelet Transform
ROC - Receiver Operating Characteristic
RNN - Recurrent Neural Networks
RFE - Recursive Feature Elimination
RAVDESS - Ryerson Audio-Visual Database of Emotional Speech and Song
SFS - Sequential Feature Selection
STFT - Short-Time Fourier Transform
SER - Speech Emotion Recognition
SB - Sub-Band
SVM - Support Vector Machine
SAVEE - Surrey Audio-Visual Expressed Emotion

TEO - Teager Energy Operator
TKEO - Teager-Kaiser Energy Operator
TEO-Auto-Env - TEO Autocorrelation Envelope Area
TEO-FM-Var - TEO-Decomposed FM Variation
TQWT - Tunable Q Wavelet Transform
VMD - Variational Mode Decomposition
WT - Wavelet Transform
ZCR - Zero Crossing Rate

CHAPTER 1

INTRODUCTION AND LITERATURE SURVEY

1.1 Background

Emotion recognition is a multidisciplinary field focused on identifying and classifying human emotions through various modalities such as speech, facial expressions, physiological signals, and text. It has become an essential area of research with applications in human computer interaction (HCI), healthcare, education, and entertainment. Speech-based emotion recognition involves extracting acoustic and linguistic features from speech signals to classify emotions like anger, joy, sadness, and fear. Facial emotion recognition, on the other hand, analyzes facial expressions from images or videos. Similarly, physiological emotion recognition relies on bio signals such as heart rate, electroencephalogram (EEG), and skin conductance to infer emotional states, while text-based emotion recognition uses natural language processing techniques to detect sentiments and emotions from textual data [1].

In most of the cases, 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 [2]. Emotion recognition systems have diverse applications, including enhancing human-computer interactions in virtual assistants and gaming, monitoring emotional well-being in healthcare, and identifying student's emotional states to enable adaptive learning environments. However, the field faces significant challenges, such as handling multilingual and multicultural variations, ensuring robust performance in noisy real world conditions, and achieving real time computational efficiency. Despite these challenges, continuous advancements in artificial intelligence and multimodal analysis are pushing the boundaries of what is achievable in emotion recognition.

In automatic emotion recognition, emotion models are divided into two categories namely discrete model and continuous model. The discrete approach focuses on basic emotions like anger, happiness, and sadness, which are seen as universal and easily distinguishable. In contrast, the continuous approach uses a multidimensional space to represent emotions, with dimensions known as emotion primitives. The most common primitives are Valence, indicating positivity or negativity, and Arousal, indicating the level of internal excitement. Some models also propose a three-dimensional space to capture the complexity of emotional states [3].

1.2 Speech Emotion Recognition

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 HCI, 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. 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 [4].

SER aims at the identification of a speaker's emotion from his or her speech. There are number of discrete emotions like happy, sad, angry, etc. in human speech, based on the different situations. The non-stationary nature of speech signal makes it difficult to predict the embedded emotion in speech signal. Apart from nonlinearity speech signal is also affected by the variation in the acoustics due to the differences in the speaking styles, variation in gender, variation in language, variety of sentences spoken, speakers with different rate of speaking. These factors straight away affect the speech features like pitch and energy that are commonly used for SER [5]. The identification of the most appropriate features for differentiating emotions is difficult. So, it is required to use the nonlinear decomposition based methods to understand the underlying pattern of speech signal. Nonlinear methods make the signal predictable and having ability to remove the noise. Decomposition also helps to target the particular area of signal for feature extraction. The choice of feature set and feature selection methods is also a challenge in SER. The speech

features are categorized as Prosodic, Spectral, nonlinear Teager Energy Operator (TEO) and Voice Quality features [6]. 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. 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 emotions i.e., anger, joy, surprise, disgust, fearful, sad and neutral [7].

1.3 Motivation

Emotion recognition significantly enhances the utility of HCI systems, making interactions between humans and machines more intuitive and effective. By detecting and analyzing emotional cues in individual’s speech, emotion recognition systems hold immense potential for monitoring and managing mental health conditions. These systems can identify signs of stress, anxiety, depression, and other emotional states, providing valuable insights that can aid in early diagnosis and intervention. This capability is particularly beneficial in the realm of mental health, where timely and accurate assessment is crucial. In healthcare, emotion recognition technology can play a transformative role. It assists healthcare providers in patient assessment by offering additional layers of information about a patient’s emotional state, which might not be apparent through traditional means. For instance, emotion recognition can help in pain management by detecting discomfort that patients may not verbally express, thereby allowing for more precise and compassionate care. Additionally, this technology facilitates communication with non-verbal or cognitively impaired individuals, who may struggle to convey their feelings and needs. By interpreting their emotional cues, caregivers can respond more appropriately, enhancing the quality of care [2].

The integration of emotion recognition into various applications leads to more personalized and responsive experiences. In virtual assistants, for example, understanding a user’s emotional state can enable the assistant to provide more empathetic and relevant responses, improving user satisfaction and engagement. Educational software can also benefit, as recognizing student’s emotions can help tailor learning experiences to keep them motivated and address any frustration or confusion they might feel. Similarly, entertainment platforms can use emotion recognition to adapt content in real-time, creating more immersive and enjoyable experiences for users. Moreover, the understanding and incorporation of emotional intelligence into artificial intelligence (AI) systems are crucial for developing ethical and empathetic AI. By recognizing and appropriately responding to human emotions, AI systems can operate in a manner that

respects and considers the emotional well-being of users [8]. This is particularly important as AI becomes more integrated into daily life, ensuring that these systems do not inadvertently cause harm or distress.

1.4 Literature Survey

In this section, existing research and literature are discussed. The presentation of the literature survey is divided into five parts. Initially, existing speech preprocessing techniques and their challenges are presented. After that, existing feature categories and feature selection techniques are discussed. Following this, performance metrics used in various studies are presented. Finally, reviews of various SER models developed using non-decomposition and decomposition-based methods are provided.

1.4.1 Speech Preprocessing

Speech preprocessing is a crucial initial step in many speech processing applications, such as speech recognition, speech-based emotion recognition, speaker identification, and speech synthesis. Fig. 1.1 shows the commonly used preprocessing method. The preprocessing stage typically includes several fundamental processes, namely framing, windowing, silence removal, decomposition and normalization. Each of these steps serves a specific purpose in preparing raw speech signals for further analysis and processing, ensuring that the subsequent stages can operate more effectively and accurately.

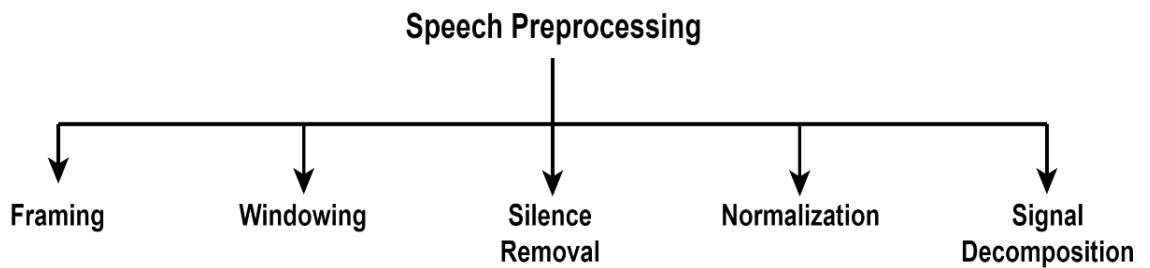


Fig. 1.1 Speech Preprocessing Methods.

Framing is the first step in speech preprocessing. It involves dividing the continuous speech signal into small segments called frames. This division is necessary because speech signals are non-stationary, meaning their statistical properties change over time. By splitting the signal into frames, we can assume each frame to be quasi-stationary, allowing for more manageable and effective analysis. Frames are typically 20-80

milliseconds long, a duration chosen because it captures the essential characteristics of speech while providing sufficient temporal resolution. Overlapping frames are often used to ensure continuity between frames and to capture transitional speech characteristics [9], [10].

Following framing, windowing is applied to each frame. Windowing involves multiplying each frame by a window function, which tapers the signal at the beginning and end of the frame. The most commonly used window functions are the Hamming, Hanning, and Blackman windows. These functions help to minimize the spectral leakage that occurs when performing a Fourier transform on the framed signal. Spectral leakage can distort the frequency components of the signal, leading to inaccuracies in further processing steps. The choice of window function and its parameters can significantly impact the quality of the preprocessing, making it a critical consideration [11].

Silence removal is another vital step in speech preprocessing. Speech signals typically contain segments of silence or non-speech sounds, such as pauses between words or phrases. These segments can introduce noise and reduce the efficiency of subsequent processing stages. Silence removal aims to eliminate these segments, focusing only on the parts of the signal that contain actual speech. This step is crucial for applications like speech recognition, where non-speech segments can negatively impact the accuracy of recognition algorithms. Techniques for silence removal often involve setting a threshold for the energy or amplitude of the signal, below which the signal is considered silent and thus discarded [12].

Normalization is the final step in the typical speech preprocessing pipeline. Normalization involves scaling the amplitude of the speech signal to a standard range. This step is essential for ensuring that variations in recording conditions, such as different microphone sensitivities or distances from the speaker, do not affect the analysis. Normalization can be done in various ways, such as peak normalization, where the maximum amplitude of the signal is scaled to a predefined value, or root mean square normalization, where the overall power of the signal is adjusted. Proper normalization ensures that the features extracted from the speech signal are consistent and comparable across different recordings [13].

The combination of these preprocessing steps forms a robust framework for preparing speech signals for further processing. Framing ensures that the non-stationary speech signal is divided into manageable segments, allowing for more effective analysis. Windowing mitigates the issues of spectral leakage, preserving the frequency characteristics of the signal. Silence removal focuses the processing on the relevant parts of the signal, enhancing the efficiency and accuracy of subsequent stages. Finally,

normalization standardizes the amplitude of the signal, ensuring consistency across different recordings [14].

The adoption of decomposition-based techniques has proven effective in boosting the performance of machine learning based architectures. Such advancements will pave the way for the practical deployment of speech emotion recognition systems in real-world applications. Several approaches based on linearity and stationary hypothesis are used for emotion detection. The most helpful method is based on Fast Fourier Transform, but this approach loses some information [15]. To overcome the stationarity issue, a short-time Fourier transform (STFT) is suggested to upgrade the conventional Fourier transform [16]. Despite improvement, STFT based approach still needs to be improved by the fundamental uncertainty. The non-stationary behavior of the signal remains challenging [17]. Hence STFT is not suitable for feature extraction from non-stationary signal. The nonlinear methods for the non-stationary call significantly improve over the linear techniques [18], [19], [20], [21].

Based on empirical mode decomposition (EMD) and wavelet transform, various methods are suggested for the statistical analysis and classification of non-stationary one-dimensional signals [22], [23]. The EMD-based decomposition method suffers from the mode mixing problem, which is solved by Huang [24] and proposes the empirical ensemble mode decomposition. Despite improvement, EMD is still noise-sensitive and limits the model's accuracy [25]. The variational mode decomposition (VMD) overcomes these limitations [25]. VMD has several advantages over EMD and wavelet transform-based methods [25]. For example, EMD is recursive, while VMD is a non-recursive method. The VMD-based approach significantly improves the statistical analysis and classification of several physiological signals [26]. VMD is also used for denoising and detecting voiced and unvoiced speech signal parts [27].

1.4.2 Feature Extraction

Speech features are crucial in developing SER system, as they capture the emotion-related information within the speech signal. These features are utilized by classification or pattern recognition models to detect emotions. Speech features are typically categorized into Prosodic, Spectral, Voice Quality, and Nonlinear TEO features [4]. Fig 1.2 shows the category and commonly used feature in each category.

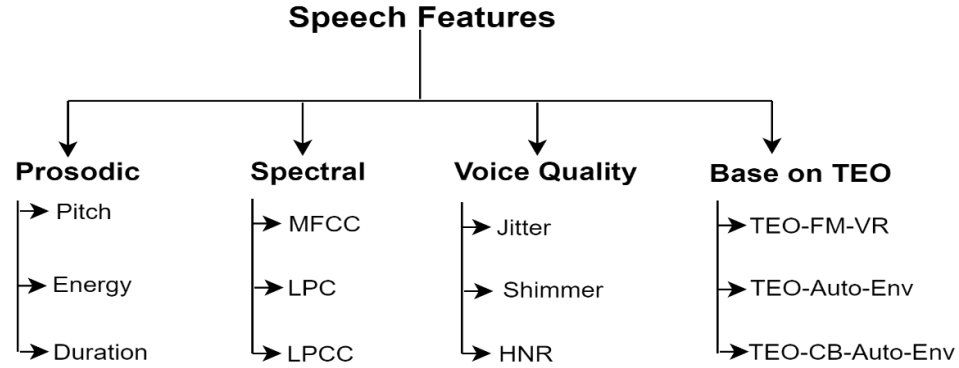


Fig. 1.2 Speech Feature Categories.

Continuous prosodic features include pitch, zero-crossing rate, energy, and formants, all of which influence the emotional tone of speech. Pitch, in particular, varies significantly with different emotions and is extensively used in SER systems to characterize emotions [28], [29]. Voice quality features, which include voice pitch, voice level, temporal and feature boundary structures, jitter, shimmer, and glottal waveforms, are strongly related to perceived emotion [30]. These features provide insight into the emotional state of the speaker through variations in the voice quality.

Spectral features represent the short-time analysis of the speech signal, capturing the spectral energy distribution which changes with emotional content. Emotions can be categorized as high-arousal (e.g., happiness, anger) or low-arousal (e.g., sadness) based on the energy levels at high frequencies. Spectral features are highly effective in characterizing emotional content compared to other speech features [31], [32] [33]. The nonlinear nature of airflow during speech production in the vocal tract system is also crucial, particularly under stressful conditions where muscle tension affects the airflow [34]. Nonlinear TEO features are employed to recognize stressed emotional speech. The TEO feature is combined with glottal features to enhance the performance of SER systems [35], [36], [37].

Identifying the optimal speech feature for emotion recognition remains a significant challenge. Mel Frequency Cepstral Coefficients (MFCC) are among the most frequently used spectral features in SER, providing promising results [38], [39], [40]. Variants like Modified MFCC (M-MFCC) and feature fusion techniques combining MFCC with Short Time Energy Features and their derivatives (velocity and acceleration) have shown improved performance over traditional MFCC and Linear Prediction Coefficients (LPC) [41]. For recognizing stressed or depressed emotions, such as anger and sadness, feature extraction techniques have evolved. These include M-MFCC, which integrates MFCC with Short Time Energy Features, enhancing SER accuracy [42]. New

approaches like sinusoidal model-based feature extraction, Empirical Mode Decomposition with feature optimization, and hybrid optimization techniques have been introduced for emotion recognition [43].

To improve SER accuracy beyond MFCC based systems, combinations of qualitative and voice quality features, weighted spectral local Hu parameters, and bio-inspired techniques like Adaptive Neuro-Fuzzy Inference Systems combined with Multi-Layer Perceptrons have been explored [44], [45]. Specific methods, such as Glottal Compensation to Zero Crossings with Maximal Teager Energy Operator, have also been proposed, though some stressed emotions remain challenging to detect accurately [46]. The TEO, introduced by Teager and Kaiser, has been used for recognizing stressed emotions, considering the auditory system's energy detection manner [47], [48]. Variants like 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) have been developed for detecting neutral versus stressed speech. Low-Level Descriptors from prosodic and spectral features, combined with TEO-CB-Auto-Env and its derivatives, have shown high accuracy in detecting clinically depressed or stressed speech. However, the complexity of combining numerous features poses challenges [49].

Feature fusion has become a common practice to enhance SER system accuracy. The 384 features are utilized in INTERSPEECH Emotion Challenge to identify emotions, out of 384 features 16 features are low-level descriptors [50], [51]. This feature set is frequently used for the SER applications. Further, this set has been expanded with additional features to create the INTERSPEECH Paralinguistic Challenge set, which includes 1582 features to improve accuracy further [52]. In summary, speech features play a critical role in SER systems, with continuous, spectral, nonlinear TEO, and voice quality features each contributing unique insights into the emotional state of speech. Continuous features like pitch and energy capture prosodic variations, while spectral features analyze frequency components. Nonlinear TEO features highlight dynamic energy changes, and voice quality features assess timbre and phonation styles. Together, these features provide a comprehensive understanding of emotions in speech. Advances in feature extraction and fusion techniques continue to enhance the accuracy and robustness of these systems, addressing the complex nature of emotion recognition in speech. Additionally, integrating domain-specific knowledge with data-driven methods helps improve feature relevance. Future research aims to develop noise-resilient and context-aware features for real-world applications. Furthermore, exploring deep learning models to automatically learn relevant features from raw speech data shows promise in reducing the dependency on handcrafted features. The incorporation of multimodal data, such as facial expressions and body language, could further enrich emotion recognition systems. Efforts to standardize emotion recognition datasets and evaluation protocols are essential for benchmarking and comparing different approaches.

1.4.3 Feature Selection

While feature fusion can enhance the classification accuracy of a Speech SER system, it also increases the computational load on the classifier. This is because not all features are beneficial for SER analysis, many features do not contribute to emotion recognition and can even degrade system performance due to the curse of dimensionality. Beyond a certain threshold, increasing the number of features leads to a decrease in SER accuracy. Thus, selecting an appropriate feature dimension is crucial for achieving optimal performance. Feature optimization or feature selection methods are essential for simplifying the task of selecting the optimal feature set [53]. This technique helps to address the dimensionality issue and also overcomes the overfitting problem by selecting a dominant feature set. A model is said to be over fitted if it functions well on training data but poorly on fresh, untested data. By reducing redundant data, feature optimization enhances emotion identification accuracy and reduces computational time and memory usage. Furthermore, the addition of more speech features increases the overfitting issue, where the model shows high accuracy during training but fails to generalize to new data. These issues can be addressed by applying feature dimension reduction techniques before classification [54], [55]. Therefore, reducing the number of features through feature selection or optimization is recommended before performing emotion classification.

Feature selection involves choosing a subset of relevant features from the original set to reduce redundancy and improve prediction performance. This process not only enhances model accuracy but also provides the cost effective model by addressing the computational complexity and storage issue [56]. Unlike feature transformation, feature selection reduces the feature space directly without transforming it. Methods such as Elastic Net, Ridge regression, recursive feature elimination (RFE), AdaBoost, and Relief-GA-ANFIS are examples of feature selection techniques. Based on the data labelling, these techniques can be divided into three categories namely semi-supervised, unsupervised, and supervised [57]. Feature selection methods are critical in optimizing models by reducing dimensionality while retaining relevant information. Supervised feature selection evaluates features using labeled data, enabling precise relevance assessment but requiring costly and time-intensive data labeling. Unsupervised feature selection operates without labels, making it ideal for large datasets, though it faces challenges in identifying feature relevance due to the lack of explicit guidance. Semi-supervised feature selection bridges the gap by combining labeled and unlabeled data, leveraging the strengths of both approaches. It maximizes utility from limited labeled data while effectively utilizing the vast unlabeled portions, making it a balanced and efficient strategy [58].

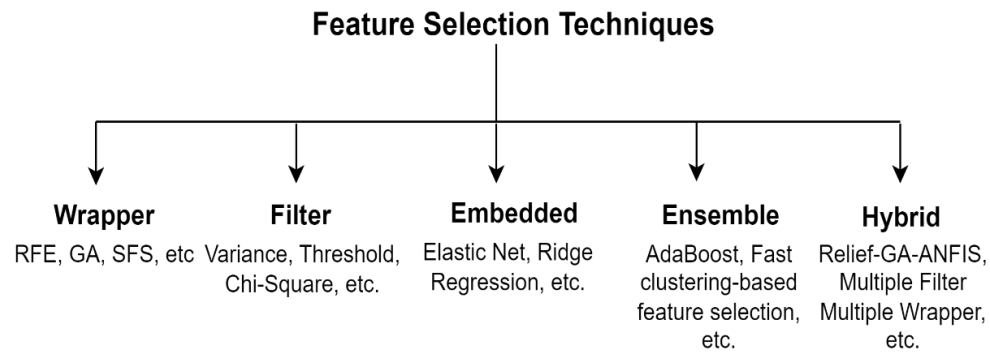


Fig. 1.3 Feature Selection Methods.

Feature selection techniques presented in Fig. 1.3 are categorized into five types: filter, wrapper, embedded, hybrid, and ensemble methods.

- **Filter Methods:** These methods are based on statistical analysis and assign scores to individual features. Features are ranked based on their scores and either retained or discarded. Common approaches in filter methods include the Chi-Square test, variance threshold, and information gain. Fisher feature selection, for instance, is employed for SER [59].
- **Wrapper Methods:** Wrapper techniques evaluate subsets of features through a search problem, assessing prediction accuracy to assign scores to each subset. This process can be systematic, stochastic, or heuristic, involving methods like genetic algorithms, RFE, and Sequential Feature Selection (SFS). SFS and Sequential Floating Feature Selection have been applied in SER [60].
- **Embedded Methods:** Embedded techniques select features during the learning process to improve model accuracy. Regularization techniques are commonly used embedded methods [61].
- **Hybrid Methods:** Hybrid approaches combine multiple feature selection methods, such as embedded and filter techniques, to maximize the benefits of each. These strategies improve efficiency, predictability, and reduce computational complexity [62].
- **Ensemble Methods:** Ensemble techniques construct collections of feature subgroups, producing aggregate results to address the unpredictability of

individual feature selection algorithms. These approaches use diverse subsampling strategies, in which a single feature selection procedure is applied to many subsamples, and the resulting features are merged for stability and robustness [63].

An example of feature selection in SER involves employing sparse representation techniques like sparse partial least squares regression [64]. Additionally, methods such as k-means clustering, sparse autoencoders, and sparse restricted Boltzmann machines can be applied for feature transformation, efficiently reducing the feature dimensions and optimizing the feature sets for SER. Techniques like adversarial autoencoders (AEs) and variational autoencoders can also encode high-dimensional feature vectors into lower dimensions while preserving the capability to reconstruct the original feature space, offering effective dimension reduction strategies for SER [65].

A novel approach involves a deep neural network-based heterogeneous model combining AE, denoising AE, and an improved shared hidden layer AE to extract features from speech signals. These layers contribute to feature optimization. To further enhance SER performance with high-dimensional feature sets, a fusion-level network with a support vector machine (SVM) classifier can be employed [66]. In summary, feature selection is vital for improving SER systems by reducing computational overhead, preventing overfitting, and enhancing model generalization. Techniques range from simple statistical methods to complex hybrid and ensemble approaches, each contributing uniquely to the optimization of feature sets for more accurate and efficient emotion recognition.

In summary, there are numerous techniques available for feature selection and optimization to reduce the dimensionality of the feature set and mitigates the drawbacks of large feature sets. Feature selection involves choosing a subset of the original features that retain the most relevant information. In machine learning, feature set of all samples is represented in an n -dimensional space known as the feature space. The feature selection or feature transformation based methods are used to reduce the dimensionality of feature space. Feature transformation changes the original feature space into a different space with a new set of axes, concentrating the discriminant feature information in a specific part of the transformed domain. On the other hand, feature selection involves picking the most significant features without transforming the original domain. Several techniques have been developed and used by researchers to select the most appropriate feature set for SER systems [67], [68]. Feature selection can boost SER system accuracy, it also increases computational complexity and the risk of overfitting. Feature selection and optimization techniques are essential to manage the feature set's dimensionality, thereby enhancing model performance and generalization while reducing computational demands.

1.4.4 Performance Matrices

In machine learning, the performance of classification tasks is often evaluated using a confusion matrix. This matrix provides a summary of the prediction results, allowing for the calculation of various performance metrics, including accuracy, recall, specificity, precision, and the receiver operating characteristic (ROC) curve. When dealing with a classification task that involves distinguishing between two or more categories, the confusion matrix is composed of four possible combinations of actual and predicted values. These combinations include true positives, true negatives, false positives, and false negatives [69].

In the context of machine learning classification, the actual value represents the true category of the target variable in the dataset, while the predicted value is the category that the model predicts for the test data. The target variable can be designated as positive or negative, often represented by binary values 1 and 0, respectively. For instance, consider a classification problem designed to predict whether a patient has a disease. Here, a positive outcome (binary '1') indicates that the patient has the disease, whereas a negative outcome (binary '0') signifies that the patient does not have the disease [70]. In the confusion matrix:

- True Positive: The model correctly predicts the positive class (e.g., correctly identifying a patient as having the disease).
- True Negative: The model correctly predicts the negative class (e.g., correctly identifying a patient as not having the disease).
- False Positive: The model incorrectly predicts the positive class (e.g., incorrectly identifying a patient as having the disease when they do not).
- False Negative: The model incorrectly predicts the negative class (e.g., incorrectly identifying a patient as not having the disease when they do).

In summary, positive and negative categories are defined by the predicted values, while the true and false categories are determined by comparing the predicted values to the actual values. In SER, the classification accuracy, which is used to assess system performance, is derived from the confusion matrix. In the context of emotion

classification, the categories correspond to different emotion classes. Accuracy provides an intuitive measure of the system's classification or prediction performance, representing the percentage of correctly identified predicted classes.

1.4.5 Overview of SER Techniques

In recent research, there are multiple machine learning and deep learning-based techniques are suggested for different SER classes. The proposed methodology based on three main stages [71]. First, it involves multi-level feature generation using Tunable Q wavelet transform (TQWT). Second, it applies the twine shuffle pattern (twine-shuf-pat) for feature extraction. Finally, discriminative feature selection is conducted using iterated neighborhood component analysis (INCA), followed by classification. This approach aims to enhance the effectiveness of feature representation and selection in the context of SER. TQWT facilitates the generation of multi-level wavelet coefficients, while twine-shuf-pat technique extracts features from decomposed wavelet coefficients. INCA feature selection method is used to select significant features. The proposed model is tested on four publically available datasets Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS), Surrey Audio-Visual Expressed Emotion (SAVEE), Berlin Emotional Database (Emo-DB), EMOtional VOice (EMOVO) and achieves 87.43%, 90.09%, 84.79%, and 79.08% classification accuracies respectively. For the mix dataset it achieves 80.05% accuracy. In another research, happy and sad emotions of RAVDESS dataset are used for the training of convolution neural network (CNN) and 66.41% classification accuracy is achieved [72].

A study presented for recognizing emotions from audio stimuli, specifically human speech, using a novel approach inspired by computer vision: the bag-of-visual-words method applied to audio spectrograms [73]. Spectrograms are treated as visual representations and analyzed with traditional computer vision techniques, such as visual vocabulary construction, speeded-up robust feature extraction, and image histogram construction. SVM classifiers are then trained. In another method bi-directional long short-term memory is used for the classification of same emotion classes happy and sad of RAVADESS dataset and achieved 70.4% recognition accuracy [74]. In a nonlinear entropy based method, the researcher aims to identify seven emotions sadness, anger, disgust, happiness, surprise, pleasantness, and neutrality using a nonlinear entropy based method [75]. Speech signals are decomposed into Intrinsic Mode Functions (IMFs), and then entropy is computed from high-frequency IMFs and averaged for mid and baseband frequencies. These entropy measures are used to generate a feature set that includes randomness features for every emotion. State of the art classifiers like Linear Discriminant Analysis (LDA), Naïve Bayes, K-Nearest Neighbor (KNN), SVM, Random Forest, and Gradient Boosting Machine are trained. Tenfold cross-validation on the Toronto

Emotional Speech dataset shows LDA achieves peak balanced accuracy (93.30%), F1 score (87.90%), and AUC (0.99) for native English speakers. The author used simple tasks, multitask feature selection/learning, and a group model for the classification of four emotions neutral, angry, happy, and sad using the RAVDESS and University of Michigan databases. The highest classification accuracy achieved for the four-class problem was 57.14% [76]. While, for the same four classes neutral, angry, happy and sad of interactive emotional dyadic motion capture database, down sampling and deep stride CNN model used and 81.70% accuracy was achieved [77]. In another method, SVM classifier along with bag of words and speeded-up robust features are utilized for five common emotion classes (sad, anger, happy, neutral and fear) classification for different datasets [73]. The maximum and minimum recognition accuracy 83% and 43% are achieved, respectively. The author explores wavelet transforms for SER, proposing a Wavelet-based Deep Emotion Recognition method using an autoencoder, 1D CNN, and Long Short-Term Memory (LSTM) networks [78]. The autoencoder reduces the dimensionality of wavelet features, and the 1D CNN-LSTM model classifies emotions. Using the RAVDESS dataset and Monte-Carlo K-fold validation, the method achieved 81.45% unweighted accuracy and 81.22% weighted accuracy in speaker-dependent experiments, outperforming state-of-the-art methods using other representations. Six different emotions neutral, disgust, anger, fear, happiness and sadness of Berlin database were classified in [79] and [80]. In first one, MFCC, pitch and energy features along with KNN and Gaussian mixture model classifier are used for the classification [79]. While in second method, MFCC features with SVM, naive bayes and KNN classifier are used. In both cases maximum accuracy 87.7% was achieved [80].

The study analyzes and classifies cold speech, which occurs when a person has a common cold, affecting the nose and throat and altering speech characteristics [81]. VMD is used to break down the speech signal into sub-signals that highlight pathological changes. From these sub-signals, statistics such as mean, peak amplitude, variance, permutation entropy, kurtosis, skewness, center frequency, energy, spectral entropy, and Renyi's entropy are extracted as features. We then use mutual information to assign a weight to these features. The proposed method outperforms traditional features like LPC, MFCC, TEO, and ComParE, achieving an emotion identification accuracy of 90.02% on the IITG cold speech dataset and 66.84% on the URTIC database. Deep belief network was introduced for the classification of six emotion classes namely surprise, neutral, anger, happiness, fear and sadness of CASIA Chinese speech emotion dataset [82] and its accuracy is analyzed and compared with the back propagation classifier [83]. In this comparison, Deep belief network and Backpropagation achieve average 92.5% and 90% recognition rate, respectively [83].

For the classification of emotions, various classification methods like Recurrent Neural Networks (RNN) [84], SVM [85], Hidden Markov Model, Neural Networks [86], and Deep learning [87] are used. Unlike shallow learning methods, a Deep

Belief Network classifier is employed for emotion recognition. The KNN model was trained using the Emo-Db and HINDI datasets, achieving a maximum accuracy of 90% for the angry class and 70-80% accuracy for the neutral class [88]. In another method, a multi-task learning approach using six binary classifiers is proposed, which achieves an accuracy of 54.76% [89]. In [90], the author employs a radial basis kernel function for four-class emotion classification, achieving a maximum recognition accuracy of 77.5%. A method based on correlation analysis and Fisher is proposed, removing redundant features with close correlations [91]. Additionally, an emotion recognition approach based on extreme learning machine decision tree is introduced to improve recognition performance further. A traditional approaches are used to develop a speech emotion recognition model for multi-language applications, with unsatisfactory recognition rates for both models [92].

A SER system using EMD and the Teager-Kaiser Energy Operator (TKEO) is proposed for efficient time-frequency analysis of non-stationary signals [93]. EMD decomposes signals into IMFs, and TKEO estimates the time-varying amplitude envelope and instantaneous frequency. Features, including novel modulation spectral and modulation frequency features based on the amplitude and frequency modulated (AM-FM) model, are extracted and combined with cepstral features to improve emotion recognition. Using SVM and RNN classifiers, the method distinguishes seven basic emotions. Experiments on the Spanish corpus achieved a 91.16% recognition rate with RNN, while on the Berlin corpus achieved 86.22% with SVM. A robust architectures for automatic SER combining with CNN and feedforward deep neural networks is proposed [94]. The models, named Backpropagation-Free Network (BFN), Categorical Neural Architecture (CAN), and Hierarchical Bayesian Network (HBN), leverage MFCC features and bag-of-acoustic-words. BFN combines bag-of-audio-words with a feedforward deep neural network, CNA is based on CNN, and HBN is a hybrid of BFN and CNA. The concatenated outputs are fed into a softmax layer to produce probability distributions for categorical classifications. A novel approach combining frame-level speech features with attention-based LSTM is proposed [95]. These features, preserving timing relations, replace traditional statistical features. Improvement strategies for LSTM, leveraging attention mechanisms to distinguish emotional saturation across frames, are introduced. However, a deep learning based approach also suffers due to the variability of speech data. Furthermore, speech data often comes in variable length sequences, and handling such variability can be challenging for deep learning-based models. While techniques like padding or masking sequences can be used, they may introduce additional complexities and affect performance. Although RNN-based architectures can capture temporal dependencies to some extent, they may struggle with capturing the complex temporal dynamics present in speech signals. This could be a limitation when dealing with subtle changes in emotional expression over time.

Most research in the field of speech emotion recognition has predominantly concentrated on a single language, with only a limited number of studies exploring MLSE. The proposed method reduces the computational complexity and is tested on the publicly available RAVDESS and Interactive Emotional Dyadic Motion Capture (IEMOCAP) [96]. The study focuses on classifying emotions from three datasets, Berlin EmoDB, IITKGP-SEHSC, and RAVDESS. Spectral features are extracted and processed, then reduced for analysis. Ensemble learning, particularly a bagged ensemble of SVM with a Gaussian kernel, is proposed for superior performance. Results are reported for the three datasets, showcasing the efficacy of the approach [97]. In another approach, author introduces two modifications for extracting MFCC [98]. In this method magnitude spectrum is used instead of the energy spectrum and omitting the discrete cosine transform, resulting in the Mel Frequency Magnitude Coefficient (MFMC). The performance of MFMC, along with three conventional spectral features was evaluated on the Berlin, RAVDESS, SAVEE, EMOVO, eNTERFACE, and Urdu databases using a multiclass support vector machine classifier. MFMC achieved accuracies of 81.50% for Berlin, 64.31% for Ravdess, 75.63% for Savee, 73.30% for EMOVO, 56.41% for eNTERFACE, and 95.25% for the Urdu database, surpassing the conventional features in performance.

A two-stream deep convolutional neural network is used to classify, and iterative neighborhood component analysis is used to optimise the feature set [99]. This technique is tested on three different databases of two other languages. The proposed framework constitutes two channels and the convolution neural network to compute the feature. The accuracy of this method is suffered by 10% with the change of language. A multi-level feature selection approach is used with uncorrelated regression for emotion classification [83]. This method is tested on four datasets of three different languages. With the change of language, the accuracy of this method is reduced by approximately 10%. A novel feature selection method is proposed to reduce computational complexity [96]. The proposed framework is used for the emotion identification from two publically available databases. A language-dependent method using a multiclass support vector machine is tested [98]. This framework achieves an accuracy of 95.25% for the Urdu language dataset and attains the maximum accuracy of 81.5% for other languages. The language dependency is also a major challenge for SER methods. A two-stream deep convolutional neural network, applied on three databases across two different languages [99]. However, a change in language results in a 10% decrease in accuracy. A multi-level feature selection method with uncorrelated regression enhances emotion identification by selecting the most relevant and independent features [71]. This approach was applied across four datasets in three different languages, showcasing its ability to generalize across linguistic contexts. However, a change in language leads to a 10% decrease in accuracy, highlighting the challenge of language-specific variations in emotional expression.

1.5 Research Gap and Problem Identification

1) Existing studies primarily focus on traditional feature extraction methods, such as spectral and prosodic features, which may not effectively capture the inherent non-stationary characteristics of emotional speech signals. The potential of advanced non-stationary decomposition techniques, such as EMD or VMD, remains underexplored for binary classification tasks in SER.

2) While significant progress has been made in multiclass emotion recognition, challenges persist due to the overlapping nature of emotional expressions in speech. Many approaches fail to generalize across multiple emotions due to suboptimal feature selection and lack of robust techniques to handle inter-class variability. The use of non-stationary decompositions to improve feature relevance and classification accuracy for multiclass SER is insufficiently addressed in the literature.

3) Many existing SER studies rely on conventional feature extraction methods and often overlook the influence of silence and noise, which may fail to capture the emotional nuances in speech signals. This limitation can affect the performance and generalizability of SER models, particularly when dealing with complex datasets that include diverse emotional expressions and gender variations.

4) Language dependence is a critical bottleneck in building scalable SER systems. The emotional cues in speech are often confounded by language-specific prosody and phonetics, making it challenging to achieve cross-language generalization. Existing research has not adequately investigated how non-stationary decomposition techniques and optimal feature sets can address the language dependency issue to create universally applicable SER systems.

1.6 Research Objectives

1) Propose a binary class speech emotion recognition method using non-stationary decompositions and optimal feature set.

2) Propose a multiclass speech emotion classification method using non-stationary decompositions and optimal feature set.

3) Propose a gender independent method for speech emotion recognition using non-stationary decompositions and optimal feature set.

4) Propose a language independent method for emotion recognition using non-stationary decompositions and optimal feature set.

1.7 SER Database

SER has evolved into a complex and challenging task within the speech processing domain. The effectiveness of an SER system in real-time environments heavily depends on the inherent nature and quality of the speech signal. Selecting an appropriate speech database for SER system development is thus crucial, as a lower-quality database can lead to inaccurate emotion classification and incorrect predictions. Several factors must be considered when developing a speech emotion database, including language, speaker gender, and number of subjects, emotion types, and age [100].

Speech corpora used in SER system development are generally categorized into three types: Actor-based (simulated), Elicited (induced), and Natural (spontaneous) emotional speech databases [100].

- **Actor-based Databases:** These databases are created using experienced and trained theatre or radio artists who perform specific emotions. These databases are fully developed and typically exhibit intense emotions, often referred to as full-blown emotional databases.
- **Elicited Emotion Databases:** These databases are recorded by simulating emotional situations without the speaker's prior knowledge. The speakers engage in emotional conversations, making these databases more natural compared to acted ones.
- **Natural Emotion Databases:** These databases contain mildly expressed emotions and are often recorded in real-life situations such as call center conversations, cockpit recordings, or dialogues between patients and doctors. Recognizing emotions in these databases can be particularly challenging due to their subtlety.

Some well-known emotion databases widely used in SER research include EMO-DB, IEMOCAP, eINTERFACE, EMOVO, SAVEE, BAUM-1s challenge database, and EMA [100]. Each of these databases has unique characteristics and is used to address various aspects of emotion recognition in speech, contributing to the ongoing development and refinement of SER systems.

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

1.7.1 RAVDESS Emotion Database

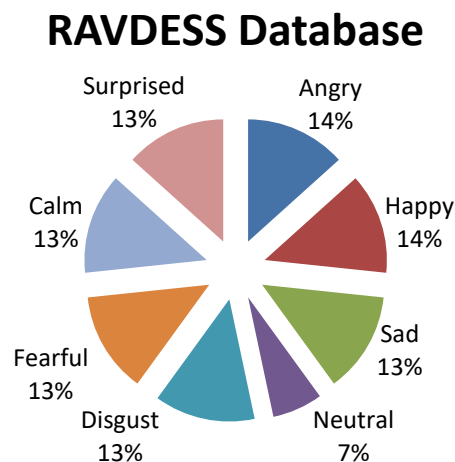


Fig. 1.4 Emotion Distribution in Percentage for RAVDESS Database.

The RAVDESS speech database was recorded in English by 24 speakers [76]. Where twelve orators are male and twelve orators are female. Every sentence is uttered twice in 2 different intensities (high and low) and eight different emotions. The available emotions are fearful, calm, sad, neutral, disgust, angry, happy, and surprised. Each emotion class database contains 192 utterances except the neutral class. In the neutral category, only 96 utterances are available. The sampling frequency of each utterance for the recording is 48KHz. The distribution of emotion class in the database is shown in Fig.1.4.

Download link: <https://zenodo.org/record/1188976>.

1.7.2 Emo-DB Database

The Emo-DB emotional speech database was recorded by ten actors in the German language [101]. Where five actors are male and five actors are female. In this database total of 535 utterances of 10 sentences are available in disgust, boredom, neutral, fear, happiness, sadness and anger. The distribution of emotion class in the database is shown in Fig. 1.5.

Download link: <http://emodb.bilderbar.info/download>

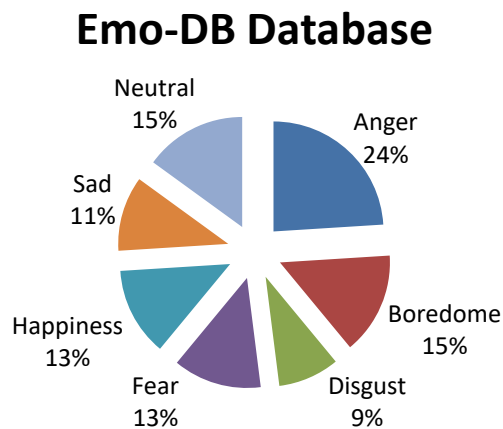


Fig. 1.5 Emotion Distribution in Percentage for Emo-DB Database.

1.7.3 IEMOCAP Database

This database is recorded in English. It is an elicited database, and for each sentence, the perceptions of different evaluators are provided [34]. For this work, the dataset is labeled based on the same perceptions held by more than half of the evaluators. In this database, measurable data are available for five emotions: happy, sad, angry, neutral, and excitement. Due to the similarity between happy and excitement, in this work, both are merged and labeled as the happy emotion, as done in other research studies [35], [36]. This process leads to a dataset of $N = 4826$ samples, with a distribution of {1108, 1336, 1159, 1223} examples for the {neutral, happy, sad, angry} classes, respectively. The distribution of emotion class in the database is shown in Fig.1.6.

Download link: https://sail.usc.edu/iemocap/iemocap_release.htm.

IEMOCAP Database

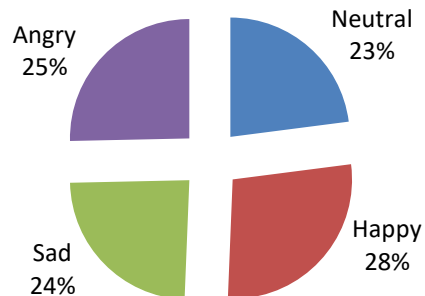


Fig. 1.6 Emotion Distribution in Percentage for IEMOCAP Database.

1.7.4 EMOVO Database

EMOVO Database

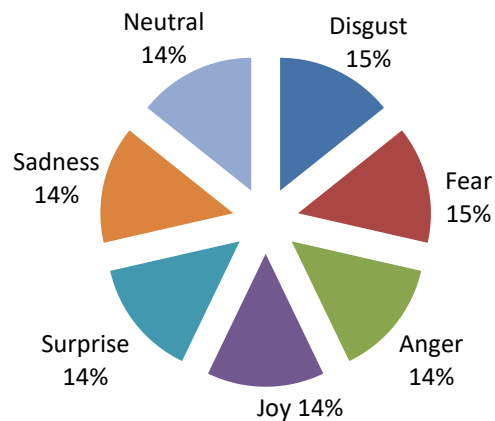


Fig. 1.7 Emotion Distribution in Percentage for EMOVO Database.

The EMOVO database was recorded by three male and three female speakers [102]. This database contains a total of 14 Italian sentences in seven different emotions. In this database total of 588 utterances are available in disgust, fear, anger, joy, surprise, sadness and neutral feelings. The sampling frequency of this database is 48KHz. The distribution of emotional utterances in the database is shown in Fig. 1.7.

Download link: <https://www.kaggle.com/datasets/sourabhy/emovo-italian-ser-dataset>.

1.7.5 Multilingual Database

The multilingual database is created using the six common emotion classes fear, happy, sad, neutral, angry, and disgust from the three publicly available emotion databases, namely the English language-based RAVDESS speech database, German language-based EMODB speech database and EMOVO Italian emotional speech database. During the mixing of database common classes are labeled with same label for example angry emotion database of all three database are combine and labeled as angry. The joy emotion of the EMOVO database is labeled as a happy emotion in the multilingual database. The distribution of emotion classes in the database is shown in Fig. 1.8. This dataset increases the number of emotions in each emotion class. Due to the mixing of different language database diversity is also created. Compare to other dataset is highly diversify and large in size.

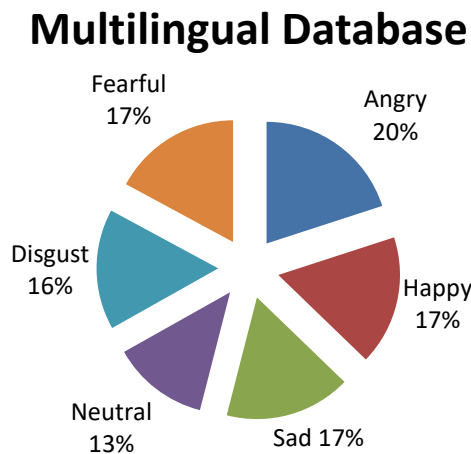


Fig. 1.8 Emotion Distribution in Percentage for Multilingual Database.

1.8 Contribution

Based on literature survey and research gap the following work has been done to accomplish proposed objectives.

1.8.1 Binary Class Speech Emotion Recognition

A rational dilation wavelet transforms (RDWT) based model is proposed for the binary emotion classification. For the features optimization KW test is performed and p value is used for the feature optimization. Finally, multiple classifiers are trained and tested. The highest accuracy is achieved with the optimizable ensemble classifier. Additionally, an empirical mode decomposition-based model is proposed for the binary emotion classification. For the features optimization rank based ReliefF algorithm is used, and high predictor importance weight are considered as significant feature. Finally, KNN classifiers is trained and tested. The proposed method proves efficient for binary class emotions.

1.8.2 Speech Emotion Recognition Using VMD-TKEO

In this method, VMD is explored and combined with Teager-Kaiser energy operator. The VMD-TKEO makes the speech signal predictable. The predictability of the proposed model is investigated for binary class and multiclass. The behavior of the proposed model is also tested for each emotion. It is observed that the proposed method is highly efficient for binary class and provides 100% accuracy. The proposed model was tested on the publicly available RAVDESS dataset and proved robust for gender-independent emotion classification applications.

1.8.3 Multiclass Speech Emotion Recognition

In this framework, a speech emotion recognition approach is presented, relying on VMD and adaptive mode selection utilizing energy information. Instead of directly analyzing speech signal this work is focused on the preprocessing of raw speech signal. Initially, given speech signal is decomposed using VMD and then the energy of each mode is calculated. Based on energy estimation, the dominant modes are selected for signal reconstruction. Following feature extraction, ReliefF algorithm is utilized for the feature optimization. The resultant feature set is utilized to train the fine K- nearest neighbor classifier for emotion identification. The predictability of preprocessed signal also checked. Which shows that the preprocessed signal are highly predictable compare to the original signal.

1.8.4 Multilingual Speech Emotion Recognition

A VMD decomposition-based framework is proposed for multilingual and gender independent emotion classification. Initially, the silence part is removed using centroid method. After that mode tuning and mode rejection is performed using Bhattacharya distance. Finally, the signal is reconstructed from the selected modes. The reconstructed signal is utilized for the prominent feature extraction and selection. The KNN classifier is trained and tested using selected feature set. The proposed model proves robust for the multilingual emotion classification application.

1.9 Organization of the Thesis

In the first chapter, basics of speech emotion recognition are presented. After that other existing state of art based on conventional method and decomposition-based method are discussed.

The second chapter presents the model for the binary emotion classification. A rational dilation wavelet transforms based model is proposed. This work proposes an emotion recognition framework using speech signal. In the proposed work speech signal is preprocessed and decomposed into SBs by RDWT. Features are extracted from RDWT provided SBs for the classification of happy and sad emotions. These features are used for the training and testing of multiple classifiers and best classifier is explored for the emotion recognition. An empirical mode decomposition based model also discussed for the binary emotion classification. In the proposed framework, the global feature extraction approach is followed. The input speech signal is decomposed into IMFs and features are computed from each IMF. In this approach two types of features are extracted, one is energy-based ratio feature and another is based on MFCC. Features are selected using the ReliefF algorithm and finally, the selected features are tested on the optimizable ensemble classifier. The obtained result is compared with the existing SER architectures in result and discussion section.

The third chapter presents binary class, three class and four class emotion classification model. The proposed framework utilizes VMD and TKEO to classify emotions using speech signals. TKEO is applied to each mode obtained from VMD, resulting in a time series signal for each mode. The features are extracted from the VMD-TKEO preprocessed signal. Features are statistically examined using the Kruskal-Wallis test. The selected features are used to train and test the SVM classifier. Finally, the proposed SER system performance is compared with the existing SER systems.

The fourth chapter presents multiclass emotion classification model. The proposed framework is focused on the preprocessing of speech signal to reduce the language dependency. The initial step involves decomposing the input signal into modes using VMD. Following decomposition, the energy of each mode is computed. Dominant modes based on energy are selected for the signal reconstruction. The reconstructed signal is utilized for the feature extraction. The potential features are chosen using a feature selection algorithm. Subsequently, these selected features are utilized to train the classifier. Finally, obtained accuracy and robustness of proposed method is compared with existing methods.

The fifth chapter presents multiclass and multilingual emotion classification model. Emotions are conveyed through subtle changes in speech parameters, such as pitch, loudness, and speaking rate. Noise in the speech signal can mask these changes, making it challenging to recognize emotions accurately. Denoising can help remove the noise and reveal subtle changes in the speech signal. Hence, this work focuses on the preprocessing of input speech data. The proposed framework has three sections. The raw speech signal is preprocessed in the first section and removes the unwanted or noisy part from the signal. The second section includes computing prosodic and spectral features from preprocessed speech signals and selecting an optimised feature set. The third section uses a fine K-Nearest Neighbors (KNN) classifier for emotion identification. The proposed architecture is tested separately on three different languages database (English, German and Spanish). Finally, the proposed architecture is tested for multilingual database.

The sixth chapter concludes the work presented in this thesis. It also provides scope and direction for future work and social impact in this field.

CHAPTER 2

BINARY CLASS SPEECH EMOTION RECOGNITION

In this chapter wavelet and EMD based approach for the binary class SER is presented. In wavelet based method, rational dilation wavelet transform, which has adjustable quality factor and high frequency resolution, is employed. Wavelet based features are extracted and statistically examined by using Kruskal-Wallis (KW) test. Further, these features are utilized for training and testing of KNN, ensemble, and decision tree classifiers variants. Among, all the classifiers optimizable ensemble classifier achieves the highest accuracy. In the EMD based approach, the proposed framework utilizes a comprehensive feature extraction method. This method involves decomposing the input speech signal into its constituent IMFs. By breaking down the speech signal in this manner, the framework can effectively capture the essential features and nuances of the input signal. The MFCC and energy-based ratio features are extracted from each IMF. The ReliefF algorithm is utilized to select dominant features, which are subsequently tested on an optimizable ensemble classifier for evaluation. Finally, the accuracy of wavelet and EMD based approach is compared.

2.1 RDWT Based SER Model

The steps involved in the proposed Rational Dilation Wavelet Transforms (RDWT) based SER model is shown in the Fig. 2.1. The workflow for SER involves several key steps. It begins with the selection of a suitable dataset, followed by pre-processing. Next, decomposition techniques are applied to isolate meaningful signal components. Feature extraction methods are then used to identify emotion relevant characteristics, which are subsequently input into classification models for emotion detection and analysis. The example of raw speech signal is shown in Fig. 2.2.

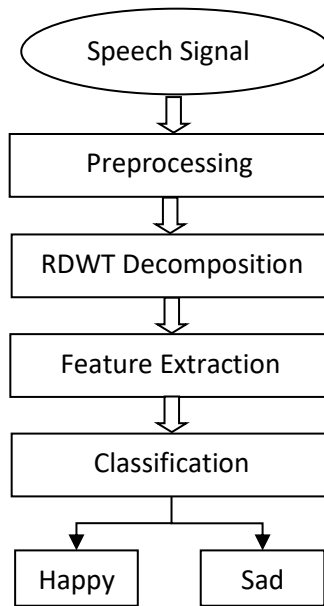


Fig. 2.1 Classification of Emotions Using RDWT Based Features.

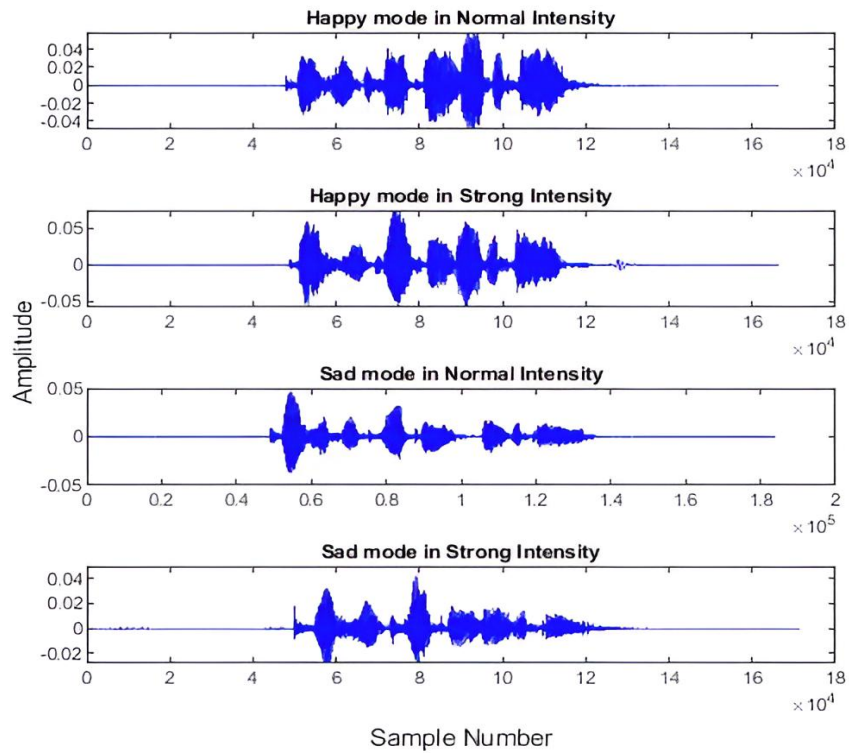


Fig. 2.2 Example of Raw Speech Signal.

2.1.1 Dataset

The proposed model uses a publicly available RAVDESS dataset [76]. The details of RAVDESS dataset is discussed in chapter 1. This work uses happy and sad emotions speech signals to form a binary classification problem.

2.1.2 Preprocessing

In this model, each speech signal is preprocessed and fragmented into 25ms duration frames. The preprocessing includes removing unvoiced part, apply Hamming Window (HW) and pre-emphasis [103]. The unvoiced part is removed by using optimize threshold value and HW is used for smoothing the edges of each frame. The example of a preprocessed signal is shown in Fig. 2.3.

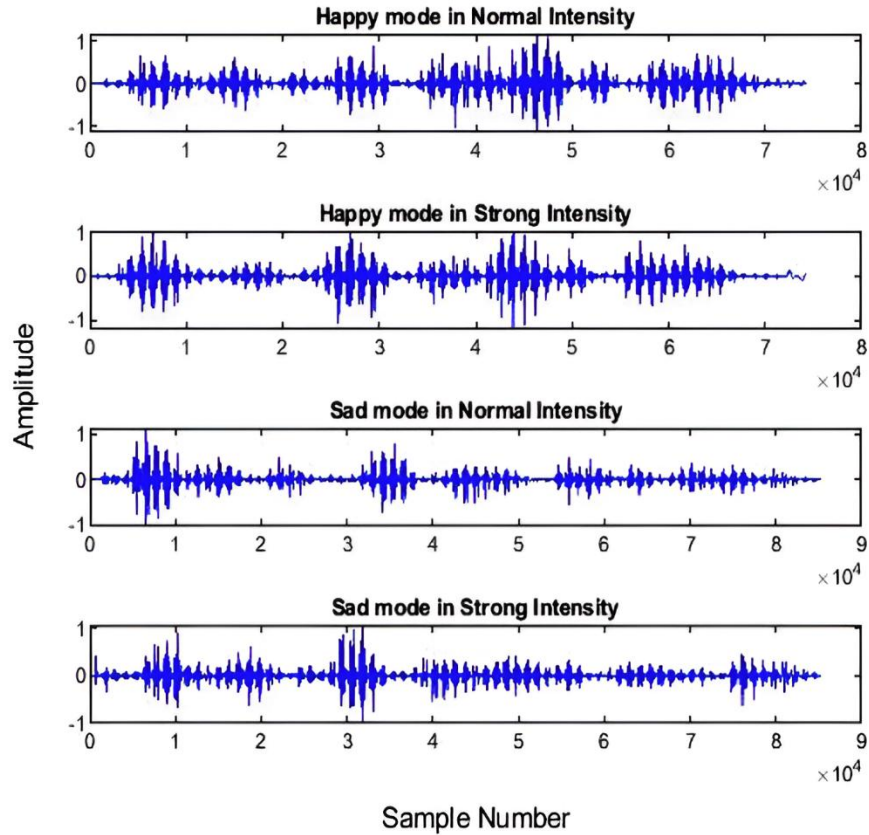


Fig. 2.3 Example of Preprocessed Speech Signal.

2.1.2.1 Rational Dilation Wavelet Transforms

In literature, many variants of wavelet transform (WT) like dual tree complex WT and double density WT are employed in signal processing. The performance of these wavelet transforms is limited by frequency resolution and low quality (Q) factor [104]. While, for the optimum representation of the oscillatory signals, a high resolution and adjustable Q factor based wavelet transform are required.

Due to the oscillatory behavior of speech signal RDWT are selected for the decomposition [104]. For designing low-pass and high-pass filters of RDWT, Fast Fourier Transform based circular convolution is used. In RDWT, the perfect reconstruction is obtained for discrete signal, when the least common multiple of p and q is multiplied with the length of signal at each level. The RDWT is approximately shift invariant and discrete transform. It provides an adjustable Q factor and high frequency resolution. For performing the decomposition RDWT uses the repeated structure of two filter banks. Further, desired Q factor is obtained by adjusting positive integer variables p , q and s . These variables must satisfy $p/q + 1/s \geq 1$ and $1 \leq p < q$ equations, where p and q are co-prime and assuring $q > p$. In most of the WT redundancy factor is fixed while in RDWT redundancy factor is controllable. The redundancy in RDWT with iterated filter-bank defined in Eqn. 2.1 [104],

$$\text{red}(p,q,s) = \lim_{j \rightarrow \infty} (p, q, s) = \frac{1}{s} \left(\frac{1}{1-p/q} \right) \quad (2.1)$$

In RDWT, the expression for scaling $\varphi(t)$ and wavelet function $\Psi(t)$ is defined in Eqn. 2.2 and Eqn. 2.3 [104],

$$\varphi(t) = \sqrt{\frac{q}{p}} \sum_n h(n) \varphi\left(\frac{q}{p}t - n\right) \quad (2.2)$$

$$\Psi(t) = \sqrt{\frac{q}{p}} \sum_n g(n) \Psi\left(\frac{q}{p}t - n\right) \quad (2.3)$$

Where, $h(n)$ and $g(n)$ represent high pass and low pass filters respectively. The mathematical value of p and q are optimized to achieve higher frequency resolution. In this work, the parameters chosen for the tuning of RDWT are $p=1$, $q=2$, $s=1$ and $j=7$. With these parameters, the preprocessed speech signal is decomposed into 8 SBs, which are further used for feature extraction.

2.1.3 Feature Extraction

In this work four basic features complexity, average amplitude change (AAC), mobility and zero crossing rate (ZCR) [105], [106] are extracted from eight sub-bands (SBs) of each frame. AAC is calculated by taking the mean of the difference between two successive samples. Mobility is measured by taking the ratio of the first derivative of variance of signal to the variance of the signal. ZCR is defined as the rate of the crossing of the zero axis by the speech signal. SBs are obtained from the RDWT decomposition of a frame of speech signal. As four features are extracted from each SB, hence there is a total of 32 effective features for a frame of speech signal. Further, these extracted features are utilized for the classification of emotions.

2.1.4 Classification Models

In this work, for the classification of emotions K-nearest neighbors (KNN), ensemble and decision tree (DT) classifier variants are tested [91]. The selection of these classifiers is based on their optimum behavior with the selected features. The ensemble method is also used to overcome the shortcoming of used classifiers. The KNN variants are Weighted KNN (WKNN), Cosine KNN (CosKNN), coarse KNN (CKNN), fine KNN (FKNN) and Medium KNN (MKNN) tested for selected feature set [88]. The DT variants are Fine tree (FT), Medium tree (MT) and Coarse tree (CoT) tested for selected feature set. The ensemble variants are boosted tree (BoT), Bagged tree (BaT), optimizable ensemble (OE), RUSboosted tree (RUSBoT) and Subspace discriminant (SubD) tested for selected feature set [5], [53].

2.1.5 Results and Discussion

In the proposed model, two emotion classes, happy and sad, from the RAVDESS dataset are used for training and testing the classifiers. Each speech signal is preprocessed and segmented into 25 ms duration frames. The preprocessing steps include removing unvoiced parts, applying a Hamming window (HW), and pre-emphasis [103]. The unvoiced parts are removed using an optimized threshold value, and HW is applied to smooth the edges of each frame. An example of a preprocessed signal is shown in Fig. 2.3. Additionally, these frames are decomposed into 8 subbands (SBs) using the RDWT. After decomposition four features complexity, AAC, mobility and ZCR are extracted from each SB, consequently, a total of 32 features are extracted from each frame. These features are statistically examined using the probabilistic (p) - value of KW test [107]. The p -value of KW test for all features is shown in Table 2.1. Most of the features are showing a

significantly low p-value for each SB. Further, these features are tested on different classifiers. For the training of classifiers, 10 fold cross validation is used.

Table 2.1 KW Test P-Values of SB-Wise Extracted Features.

| SB/ Feature | Complexity | AAC | Mobility | ZCR |
|------------------------|-------------------------|------------------------|-------------------------|-------------------------|
| SB-1 | 1.04×10^{-120} | 1.23×10^{-18} | 3.24×10^{-12} | 1.65×10^{-62} |
| SB-2 | 8.15×10^{-22} | 2.91×10^{-31} | 1.18×10^{-36} | 3.35×10^{-43} |
| SB-3 | 0.01 | 2.27 | 0.01 | 0.19 |
| SB-4 | 0.9571 | 4.39×10^{-18} | 8.09×10^{-68} | 1.10×10^{-82} |
| SB-5 | 3.77×10^{-22} | 0.68 | 1.16×10^{-117} | 4.40×10^{-117} |
| SB-6 | 5.60×10^{-8} | 0.17 | 0.35 | 0.37 |
| SB-7 | 3.83×10^{-29} | 3.05×10^{-6} | 4.74×10^{-13} | 9.17×10^{-19} |
| SB-8 | 2.42×10^{-98} | 1.16×10^{-13} | 1.79×10^{-19} | 1.64×10^{-19} |

Table 2.2 Proposed Methods Classification Accuracies with DT, KNN and Ensemble Classifiers Variants.

| DT | | KNN | | Ensemble | |
|-----------------------|-------------------------|------------------------|-------------------------|-----------------------------|-------------------------|
| DT Variant | Accuracy (%) | KNN Variant | Accuracy (%) | Ensemble Variant | Accuracy (%) |
| FT | 67.4 | FKNN | 79.9 | BoT | 70.2 |
| MT | 63.3 | MKNN | 76.9 | BaT | 78.9 |
| CoT | 60 | CKNN | 72.2 | SubD | 52.6 |
| | | CosKNN | 76.9 | RUSBoT | 65 |
| | | WKNN | 79.1 | OE | 83.3 |

By observing Table 2.2, for speech emotion classification KNN and ensemble classifier variants exhibit better performance in comparison to DT classifier variants. In KNN variants, FKNN achieves the highest 79.9% and CKNN lowest 72.2% classification accuracy. Similarly, in ensemble variants, BoT achieves the lowest 70.2% and OE highest 83.3% classification accuracy. The classification accuracy of other variants such as FT, MT, CoT, MKNN, CosKNN, WKNN, BaT, SubD and RUSBoT is also shown in Table 2.2. Among all the classifiers OE classifier achieves the highest accuracy 83.3% for the selected classes (happy and sad) and dataset (RAVDESS). Fig. 2.4 (Where, -1-Happy and 1- Sad) represents the confusion matrix of the OE classifier. The truly classified signal in

one class is 15710 and in the other class is 13967. The performance of the OE classifier is further evaluated through the ROC curve in Fig. 2.5. In this figure, it can be observed that the area under curve (AUC=0.92) is very high. The high value of AUC represents the stable classification of happy and sad class signals.

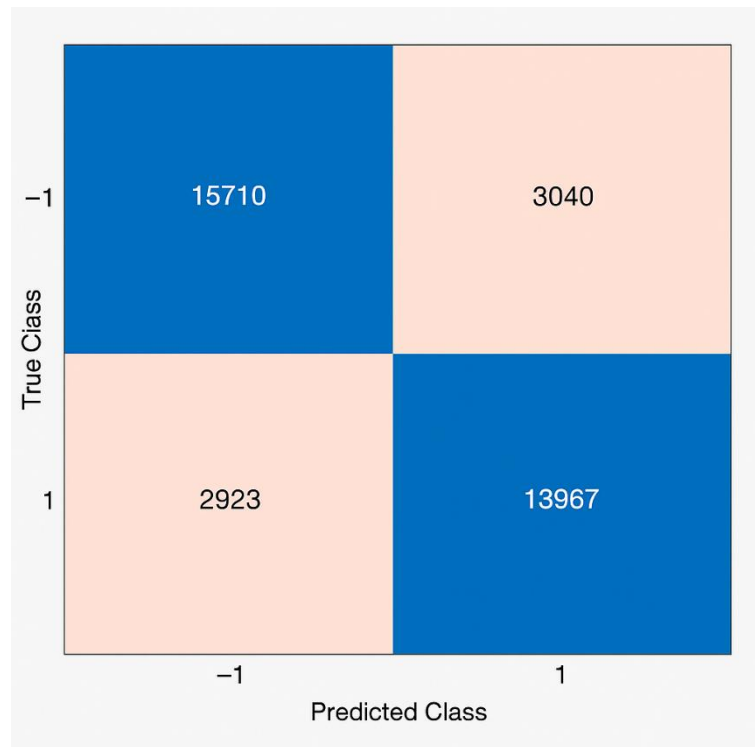


Fig. 2.4 Confusion Matrix of OE Classifier.

Table 2.3 Performance Comparison with Existing State of Art.

| Authors | Classified emotions | Classifier | Dataset | Accuracy (%) |
|-------------------------|---------------------|------------|----------------|--------------|
| R. Jannat et al. [72] | Happy, Sad | CNN | RAVDESS | 66.41 |
| M. A. Jalal et al. [74] | Happy, Sad | BLSTM | RAVDESS | 70.14 |
| Proposed Method | Happy, Sad | OE | RAVDESS | 83.30 |

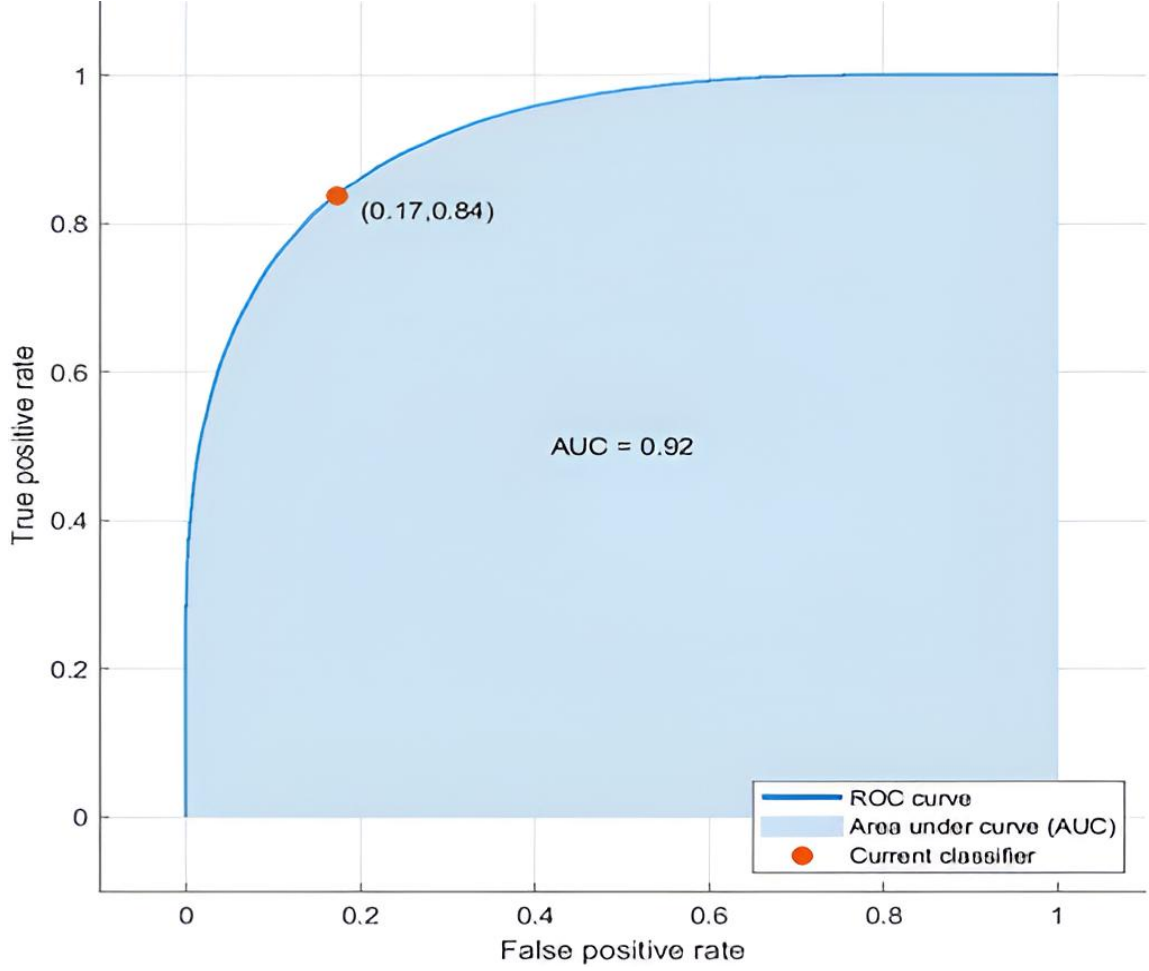


Fig. 2.5 ROC Curve of OE Classifier.

RAVDESS dataset recorded in high and low intensity and also includes both male and female speakers. Hence there are very few methods available, which include all variations. In Table 2.3, the performance of the pre-established method on the same RAVDESS dataset is compared with the proposed methods. The classification accuracies achieved in previous methods are 66.41% by R. Jannat et al [72] and 70.14% by M. A. Jalal et al [74], while the proposed method achieves 83.30% classification accuracy. So, this work has proposed more accurate method for emotion recognition using speech signals. Furthermore, the proposed method demonstrates significant improvements in handling variations in speech quality and background noise, outperforming previous methods in robustness. The comparison also highlights the enhanced generalizability of the proposed method across different speech emotions, achieving consistent results across multiple test sets.

2.2 SER Model based on Empirical Mode Transform

The process flow of proposed SER model is presented in Fig. 2.6. The detail description of each stage of Fig. 2.6 is discussed in subsequent sections.

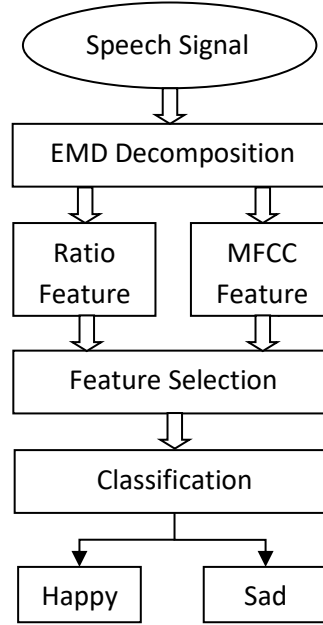


Fig. 2.6 Process Flow of Proposed SER Model.

2.2.1 Dataset

In the proposed framework publically available EMOVO and RAVDESS dataset is used [102]. The details of EMOVO and RAVDESS dataset is discussed in chapter 1. In this experiment, the emotions joy or happy and sad are used for testing the proposed method. The selection of these emotions is based on their distinct acoustic features, which are easier to differentiate in speech signals. Additionally, the EMOVO dataset provides a rich set of emotional speech samples from native Italian speakers, while the RAVDESS dataset offers a diverse set of emotions with a broader demographic representation. The use of both datasets ensures the robustness and generalizability of the proposed method. Future experiments may incorporate more complex emotions, such as anger or surprise, to further assess the performance across a wider emotional spectrum.

2.2.2 EMD Decomposition

The input speech signal initially decomposed using EMD, which dissects the non-stationary speech signal into an AM-FM oscillating components termed as IMFs [23]. An IMF is a function that adheres to the following criteria:

- Either the number of extrema is the same or it changes by a maximum of one.
- Based on local extrema, the mean value of envelopes is zero.

The EMD procedure for extracting IMFs from a time series signal $x(t)$ can be delineated through the underneath steps:

1. Determine each signal's $x(t)$ local maximum and local minimum.
2. To obtain the envelopes $h_{max}(t)$ and $h_{min}(t)$, connect all the maxima and minima individually.
3. Compute the average value of the envelopes using Eqn. 2.4 [23]:

$$m = \frac{h_{max}(t) + h_{min}(t)}{2} \quad (2.4)$$

4. Deduct $m(t)$ from the $x(t)$ as follows [23]:

$$q_1(t) = x(t) - m(t) \quad (2.5)$$

5. Verify whether $q_1(t)$ complies with the specified Intrinsic Mode Function (IMF) conditions mentioned earlier.
6. Continue from steps 2 through 5 until an IMF is obtained.

Once the initial IMF is acquired, define $I_1(t) = q_1(t)$, representing the smallest temporal scale in $x(t)$. The subsequent IMF can be obtained by generating a

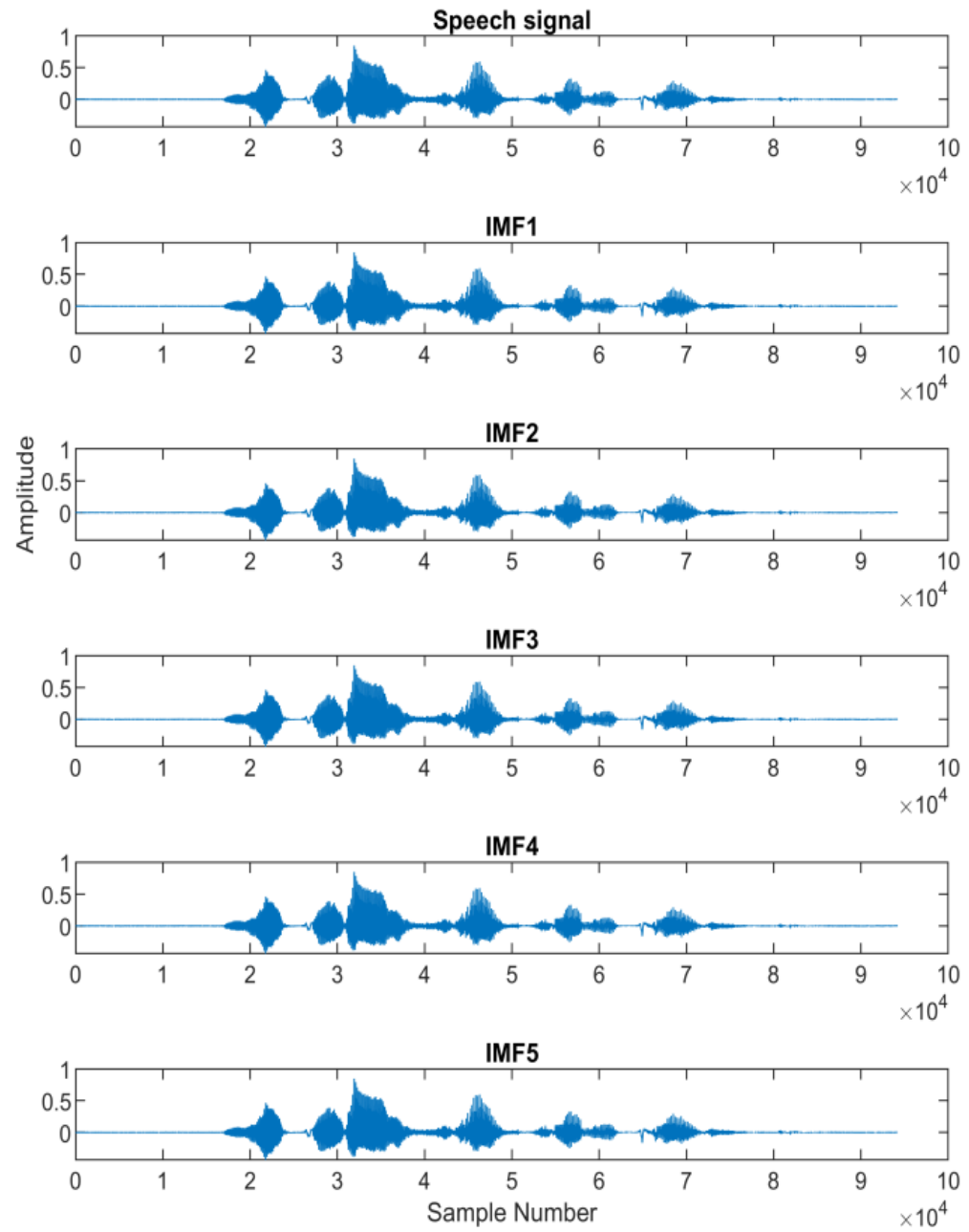


Fig. 2.7 Speech Signal and IMFs after EMD Decomposition.

residue $r_1(t) = x(t) - I_1(t)$, which is then used as the new signal in the aforementioned algorithm. The obtained residue is iterated until it becomes a monotonic function, signifying that no further IMFs can be produced. A set of narrow-band symmetric waveforms is formed by the resulting IMFs. The Fig. 2.7 shows the speech signal and their IMFs after decomposition. Following the decomposition, the signal $x(t)$ can be defined as in Eqn. 2.6 [23],

$$x(t) = \sum_{k=1}^K I_k(t) + r_K(t) \quad (2.6)$$

Where $I_k(t)$ is k_{th} IMF, K represents total number of IMF and $r_K(t)$ is the final residue.

2.2.3 Feature Extraction

In the proposed framework ratio feature based on energy and statistical measures are calculated from MFCC coefficients. The detail descriptions are available in following sections.

2.2.3.1 Ratio Feature

After decomposition the energy of each IMF is calculated using Eqn. 2.7 [108],

$$E_k = \sum_{n=1}^N |I_k(t)|^2 \quad (2.7)$$

Where, N is preprocessed signal length.

To obtain the ratio feature each IMF, feature energy is divided by the energy of remaining IMFs. The ratio features from IMF is calculated using Eqn. 2.8,

$$f = E_k / E_l \quad l = 1, 2, 3 \dots \dots \text{except } k \quad (2.8)$$

2.2.3.2 MFCC Feature

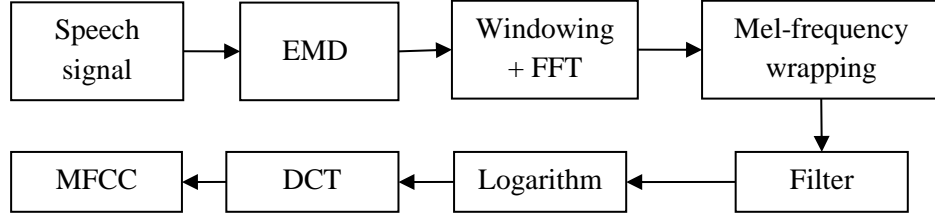


Fig. 2.8 Schema of MFCC Feature Extraction.

The MFCC serves as a prominent cepstral feature in the development of SER systems. Its primary purpose is to accurately depict the short-time power spectrum of a speech signal. Each preprocessed signal undergoes MFCC extraction, as outlined in Fig. 2.8.

To extract MFCC, the IMF obtained from EMD undergo a series of steps, including windowing (with a frame duration of 30ms and 20ms overlapping), performing the discrete cosine transform after mel-frequency wrapping and periodogram calculation [109]. The Fast Fourier transform, which converts the time-domain signal into the frequency domain, is computed using the discrete Fourier transform (DFT). Applying DFT to IMF transforms each time domain frame of N samples to frequency domain. The definitions of DFT and the Mel scale are provided in [109],

$$Y_p = \sum_{n=0}^{N-1} s(n) e^{-\frac{j2\pi hn}{N}} \quad p = 0, 1, 2, 3 \dots \dots N - 1. \quad (2.9)$$

Where, $s(n)$ is IMF obtained after applying EMD on speech signal and N is frame length,

$$M(f) = 1125 \ln\left(1 + \frac{f}{700}\right) \quad (2.10)$$

Where, f is speech signal frequency.

2.2.4 Feature Selection and Classification

The feature selection method employs the ReliefF algorithm, a rank-based approach, to choose relevant features [110]. This algorithm randomly selects instances from both the same class and different classes. When R data instances are randomly chosen from the total n instances, the ReliefF algorithm computes the feature score for each original feature as discussed in [110].

In proposed framework optimizable ensemble classifier is used for the emotion identification. It combines multiple individual models, such as decision trees or neural networks, to enhance predictive performance by leveraging diverse perspectives [111]. Through optimization techniques like boosting or bagging, it strategically weighs and combines the strengths of each constituent model, effectively mitigating individual weaknesses. This approach promotes robustness, generalization, and improved accuracy in complex classification tasks, making it a powerful tool for heightened performance and adaptability across dataset.

2.2.5 Results and Discussion

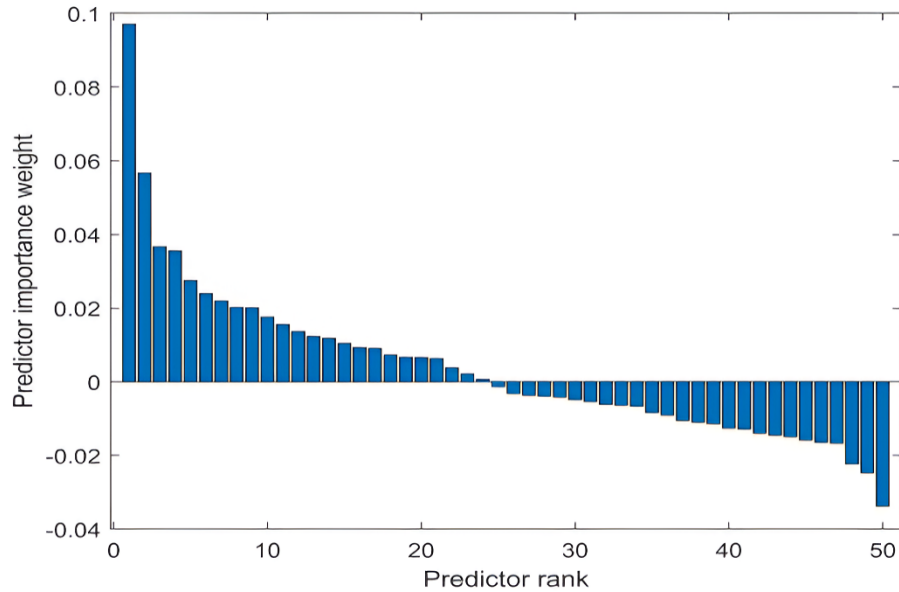


Fig. 2.9 Relieff Algorithm Predictor Rank Bar for Feature Selection.

In this experiment publically available EMOVO and RAVDESS database is used to train the classifier. Initially, the speech signal is decomposed into five IMFs using EMD method. After that energy based ratio feature and MFCC based statistical features are calculated. For the energy based ratio features, the energy of each IMF is calculated and obtained energy of each IMF is divided with remaining IMF energy as explained in section 2.2.2.

Table 2.4 Precision, Recall and F1 Score of Proposed Model for EMOVO Dataset.

| Emotions | Precision (%) | Recall (%) | F1 score (%) |
|----------|---------------|------------|--------------|
| Joy | 94.25 | 96.47 | 95.34 |
| Sad | 96.38 | 94.11 | 95.23 |

Table 2.5 Precision, Recall and F1 Score of Proposed Model for RAVDESS dataset.

| Emotions | Precision (%) | Recall (%) | F1 score (%) |
|----------|---------------|------------|--------------|
| Joy | 89.25 | 90.75 | 89.97 |
| Sad | 91.35 | 88.60 | 89.95 |

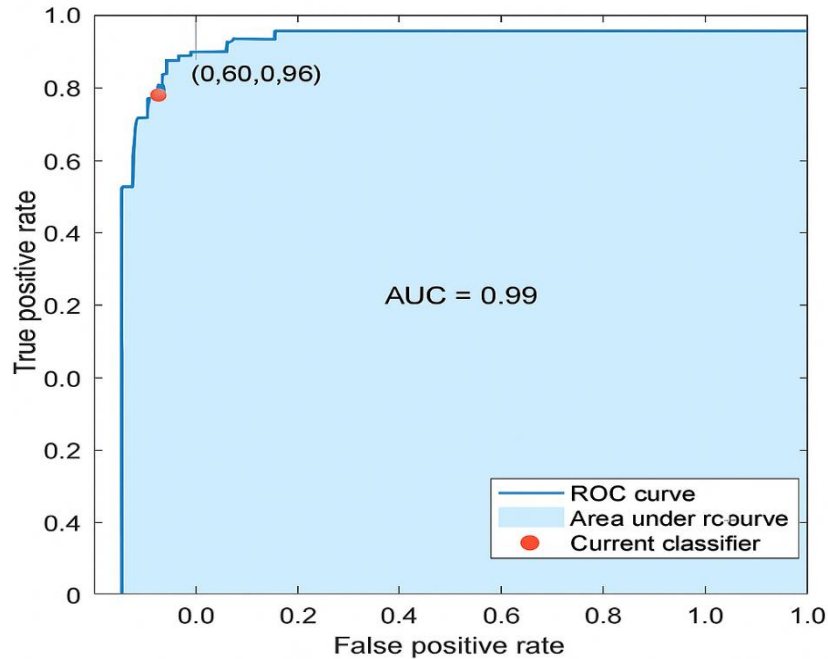


Fig. 2.10 Region of Convergence of Proposed Framework.

In MFCC based features mean and variance are calculated using the first three coefficients because most of the information is lies in this region. The extracted features are optimized using ReliefF algorithm. Initially a total of 50 features are extracted. After that, top 25 features having high predictor importance weight are considered as significant feature and remaining are rejected. The predictor rank bar is shown in Fig. 2.9.

These selected features are utilized to train the classifier. The classifier is trained using 10 cross validation in which complete dataset is divided in 10 subsets. In which 9 subsets are utilized for the training and remaining 1 is used for testing. This process is continuing until all subsets have been utilized as the test dataset. The OE classifier is used for emotion classification and achieves an impressive 95.3% recognition accuracy. Additionally, Table 2.4 and Table 2.5 presents performance indicators for each emotion, such as recall, precision, and F1 score. Notably, the F1 scores show that the proposed framework exhibits a similar degree of efficacy for both emotions. Further, Table 2.4 and Table 2.5 show how robust the framework's design is, since it constantly produces reliable outcomes for both emotional states. The precision and recall demonstrate how effectively the framework can recognize and categorize emotions. Finally, the proposed framework's effectiveness is confirmed by the high recognition accuracy and balanced F1 scores across emotions, shows its potential for use in a variety of situations where emotion classification is required. This work advances the field by establishing the framework as a reliable and strong solution for emotion recognition applications. Fig. 2.10 shows the region of convergence of proposed framework.

Table 2.6 Comparison of Proposed Framework with Existing SER Model.

| Authors | Classified Emotions | Method | Dataset | Accuracy (%) |
|-------------------------|----------------------------|--------------------------------------|----------------|---------------------|
| R. Jannat et al. [72] | Happy, Sad | Ubiquitous computing with CNN | RAVDESS | 66.41 |
| M. A. Jalal et al. [74] | Happy, Sad | Used BLSTM, CNN and Capsule networks | RAVDESS | 70.14 |
| Proposed Method | Joy, Sad | EMD with OE classifier | EMOVO | 95.3 |
| Proposed Method | Happy, Sad | EMD with OE classifier | RAVDESS | 90.01 |

The comparison between the suggested framework and the current SER models is presented in Table 2.6. Earlier research showed that M. A. Jalal et al. [74]

obtained 70.14% classification accuracy, R. Jannat et al. [72] obtained 66.41%. In contrast, the proposed method in this study attains a significantly higher classification accuracy of 95.3% for EMOVO and 90.01% for RAVDESS dataset. Therefore, this research introduces a more precise approach for emotion recognition utilizing speech signals, surpassing the accuracies achieved by previous methodologies. The substantial improvement in classification accuracy underscores the effectiveness and advancement offered by the proposed method in accurately discerning emotional states from speech signals. This enhancement in performance positions the proposed method as a noteworthy contribution to the field of emotion recognition, showcasing its potential to outperform existing approaches and enhance the reliability of emotion classification systems. Moreover, the proposed framework demonstrates superior performance across various acoustic conditions, proving its robustness in real-world applications. The integration of advanced feature extraction techniques further boosts the system's ability to capture subtle emotional cues in speech. Finally, this improvement in accuracy paves the way for future developments in emotion-aware technologies.

2.3 Comparison of RDWT and EMD based SER Model

The Table 2.7 presents a comparison of the accuracy percentages of two proposed methods. The RDWT-based approach achieves an accuracy of 83.3%, whereas the EMD-based approach significantly outperforms it with an accuracy of 90.01%. The results shows that the EMD based approach performed much batter compared to the wavelet based method. The higher accuracy of the EMD-based approach suggests it is better suited for achieving reliable and precise results, highlighting its potential advantages over the RDWT-based method in practical applications. This comparison underscores the importance of method selection in optimizing classification performance.

Table 2.7 Performance Comparisons of Proposed Approaches.

| Proposed method | Emotions | Dataset | Accuracy (%) |
|---------------------|------------|---------|--------------|
| RDWT based approach | Happy, Sad | RAVDESS | 83.30 |
| EMD based approach | Happy, Sad | RAVDESS | 90.01 |

2.4 Summary

This chapter presents a comparative analysis of two methodologies for emotion recognition in speech signals. In first approach RDWT decomposes speech frame into 8-SBs. Further, four features complexity, AAC, mobility and ZCR are extracted from

each SB and their statistical significance is examined by using the p -value of KW test. These features are utilized for the training and testing of multiple classifiers variants. It is observed that KNN and ensemble variants exhibit better performance in comparison to DT variants. Among all the classifiers OE classifier exhibits the highest accuracy 83.3%, which is also highest as compared to existing works. In second method, EMD decomposes input speech signals into IMFs and computes features such as ratio and MFCC from each IMF. The statistical significance of these features is evaluated using the ReliefF algorithm, and the highest-ranking features are selected for further analysis. The energy-based ratio feature shows significant potential for classification. These features are used to train OE classifiers with 10-fold cross-validation, achieving an impressive accuracy of 95.3% for EMOVO dataset and 90.01% for RAVDESS dataset. This method is also compared with existing state-of-the-art techniques and is found to be more efficient and robust. The comparison reveals that the data driven EMD based approach significantly outperforms the RDWT based approach in terms of accuracy, with the former achieving 90.01% and the latter 83.3%. The superior performance of the EMD based method shows that the data driven method is more effective for the SER in comparison to wavelet based methods.

CHAPTER 3

SPEECH EMOTION RECOGNITION USING VARIATIONAL MODE DECOMPOSITION

In previous chapter, it was observed that the data-driven EMD method performed better compared to the wavelet-based approach. Therefore, in this chapter, another data-driven method, VMD, is tested. VMD was chosen because it addresses the limitations of EMD. The proposed method explores VMD with the TKEO for the SER. First, VMD decomposes a speech signal into modes, and then the nonlinear TKEO operator is applied to each mode to obtain a time series. The VMD-TKEO preprocessed signal is used to extract the global features based on Energy, Pitch frequency and Mel frequency cepstral coefficients. The features are statistically examined using the KW test. The resultant feature set is examined over the support vector machine and its variants for emotion classification. The RAVDESS speech database is used for the experiment, and different emotion classification problems are formulated. Finally, the accuracy of the proposed SER architecture is quantitatively analyzed, which outperforms the other existing architectures. To demonstrate the superiority of the VMD-TKEO method, it is compared with the EMD-TKEO method.

3.1 SER Model Using VMD-TKEO

In the proposed VMD-TKEO and EMD-TKEO algorithm, VMD or EMD decomposes a signal into its oscillatory modes, while TKEO is used to calculate the instantaneous energy of the signal. The primary advantage of the VMD-TKEO method is its precision in decomposing a signal into its oscillatory modes and determining the instantaneous energy of each mode. Additionally, the VMD-TKEO method can effectively analyze signals with non-uniform frequency content, which traditional signal processing techniques may struggle to do. Moreover, it is a data-driven method that does not require prior knowledge of the signal or its properties, offering a further advantage.

The complete block diagram representation of proposed model is presented in Fig. 3.1 and Fig. 3.2. Before feature extraction from raw speech signal following preprocessing is done (i) VMD or EMD decompose a speech signal into modes or IMFs (ii) Apply the TKEO to each mode to obtain a simplified preprocessed signal. Features are extracted from each preprocessed signal and optimized using KW test. Finally, the selected features are utilized to train and test machine learning models. The detailed explanation of proposed framework is presented in the following sections.

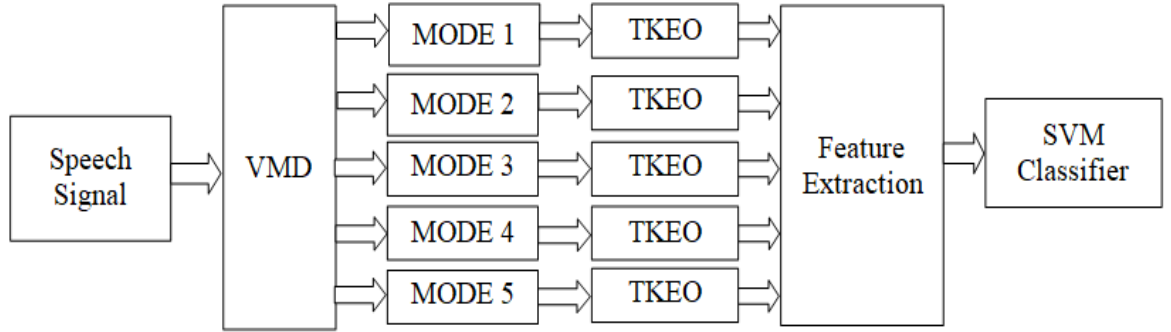


Fig. 3.1 Overall Block Diagram of Proposed SER Architecture Based on VMD.

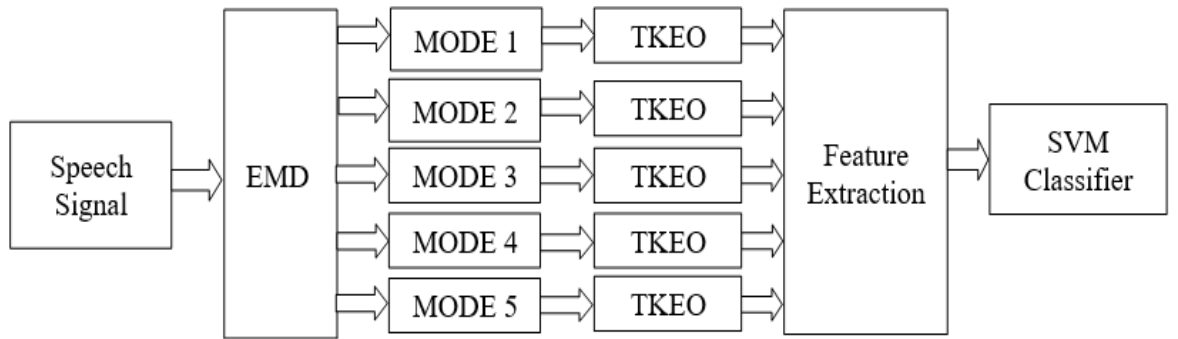


Fig. 3.2 Overall Block Diagram of Proposed SER Architecture Based on EMD.

3.1.1 Variational Mode Decomposition

The initial step of the proposed model is to decompose the raw speech signal x into K number of modes using the VMD method. VMD is a signal processing technique that decomposes a signal into oscillatory modes, each with a specific frequency and amplitude. These oscillatory modes are called SBs and are arranged in order of increasing

frequency. The modes with lower frequencies are considered to have a larger scale and slower variation, while the modes with higher frequencies have a smaller scale and faster variation. Therefore, there is a hierarchy in the oscillation of different modes, where the lower-frequency mode provide a coarse signal approximation. In contrast, the higher-frequency mode provide more detailed information about the signal [25]. The modes of the decomposed signal must possess the given two properties:

- 1) In a complete speech signal, the number of extrema must either be equal or differ by one.
- 2) The envelope defined by the local extrema should have an average value of zero at all times.

The fundamental principle of the VMD technique involves decomposing non-stationary signals into modes, where each mode is centered around a specific frequency, s_k , that is estimated during the decomposition process [25]. The amplitude and frequency of each mode at any given point in time are determined using the Hilbert transform, which generates a unilateral spectrum. To calculate the bandwidth of each mode, the following steps are taken:

1. Obtain the unilateral spectrum of each mode using the Hilbert transform.
2. Shift the unilateral spectrum towards the baseband by mixing with an exponential that is tuned to the center frequency.
3. Determine the signal bandwidth using Gaussian smoothness. This process involves solving an optimization problem, which is defined as in Eqn. 3.1 and Eqn. 3.2 [25],

$$\min_{u_k, s_k} \left\{ \sum_{k=1}^K \left\| \partial_n \left[\left(\delta(n) + \frac{j}{\pi n} \right) * d_k(n) \right] e^{-j\omega_k n} \right\|_2^2 \right\} \quad (3.1)$$

Subject to,

$$\sum_{k=1}^K d_k(n) = x \quad (3.2)$$

Where s_k is central frequency [34], n represents the discrete time index and $d_k(n)$ represent k^{th} mode of a signal. In this work, the number of modes is five with an initial ω value of 1.

3.1.2 Teager-Kaiser Energy Operator

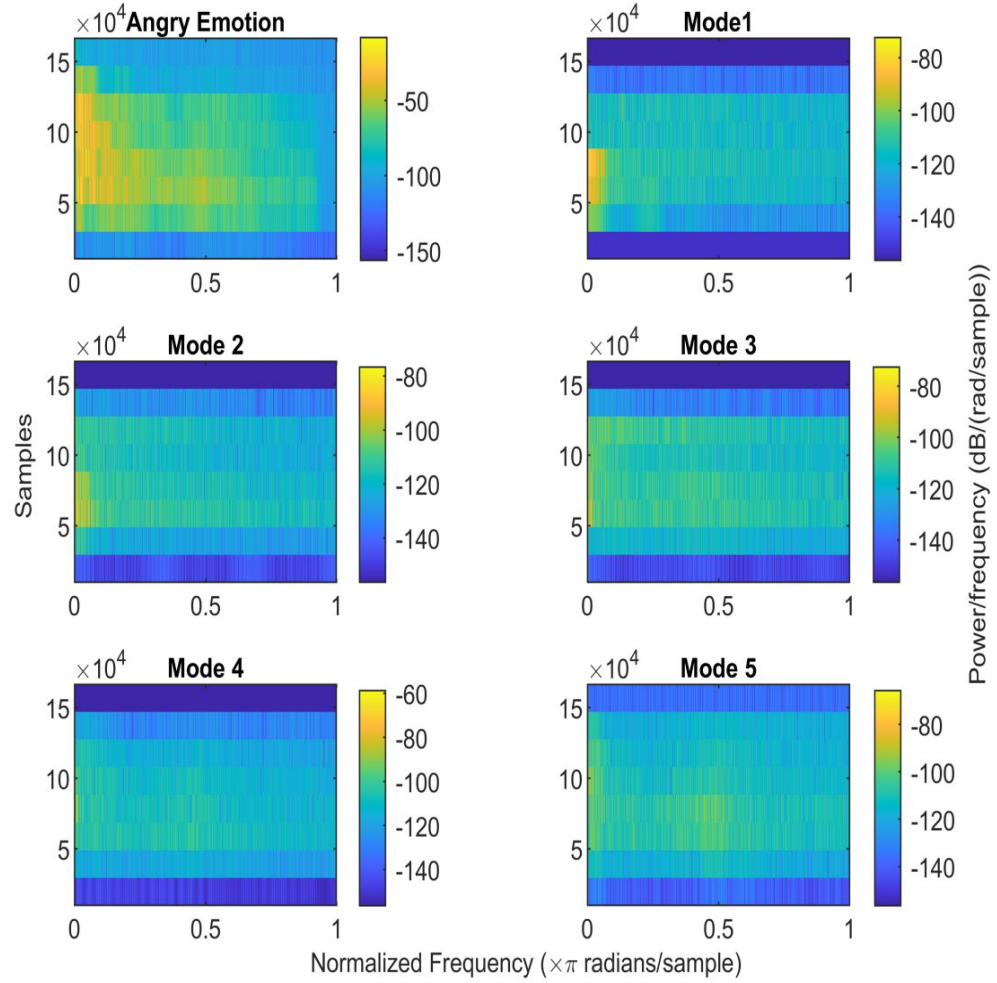


Fig. 3.3 Spectrogram of Raw Speech Signal in Angry Emotion and VMD-TKEO Preprocessed Signal.

The raw speech signals are decomposed using VMD into modes. Still, modes obtained from decomposition need to be explored more for the physical characteristics

extraction of a speech signal. This problem is overcome by using the TKEO operator. TKEO is a nonlinear energy operator that estimates mode's instantaneous energy. TKEO processed signal for a mode $d(n)$ can be defined as in Eqn. 3.3 [93]:

$$y(n) = d^2(n) - d(n+1)d(n-1) \quad (3.3)$$

where $y(n)$ is the VMD-TKEO processed signal.

VMD combined with TKEO gives the best time series analysis of the non-stationary signals. Fig. 3.3 shows the spectrogram of the raw speech signal in angry emotion and all five decomposed signals after applying VMD-TKEO.

3.1.3 Feature Extraction

The proposed algorithm employs a global feature extraction approach by computing features from VMD-TKEO and EMD-TKEO preprocessed signals. Specifically, energy, pitch frequency, and MFCC-based features are extracted for emotion classification. The energy of each VMD-TKEO and EMD-TKEO preprocessed signal is calculated directly, capturing the overall signal power, which is critical for detecting emotional intensity. Meanwhile, the statistical measures (such as mean and variance) of MFCC and pitch frequency are computed after framing the preprocessed signal to account for temporal variations in speech. This process ensures that both spectral and prosodic information, essential for distinguishing emotions, are effectively captured. A more comprehensive explanation of the feature extraction process is provided in the subsequent subsections.

3.1.3.1 Energy

In this work, energy is calculated for each preprocessed signals. It is one of the basic feature, which plays an important role for the classification of high energy emotions like angry, happy and low energy like calm emotions [108]. The energy of each preprocessed signals (signal length N) is calculated using Eqn. 3.4,

$$E = \sum_{n=1}^N |y(n)|^2 \quad (3.4)$$

3.1.3.2 Pitch Frequency

The preprocessed signal $y(n)$ obtained from VMD-TKEO is further segmented into frames (the duration of frame is 52ms and overlapping is equal to 42ms) and the pitch frequency for each frame is computed using the autocorrelation-based technique [4]. The autocorrelation of a frame $c(n)$ having length N_l can be obtained using Eqn. 3.5,

$$Aut_l = \frac{1}{N_l} \sum_{n=1}^{N_l-1} c(n) c(n + l) \quad (3.5)$$

Where, l is shifting parameter and Aut_l is autocorrelation function.

Now, the mean and variance statistical measures are calculated based on the pitch frequency obtained from all the frames to assess the global characteristics of the preprocessed signal [4].

3.1.3.3 Mel Frequency Cepstrum Coefficients

MFCC is widely used cepstral feature for the designing of the SER system. It is used for the correct representation of the short time power spectrum of an audio signal. MFCC is extracted from each preprocessed signal. The process for the MFCC extraction is presented in Fig. 3.4.

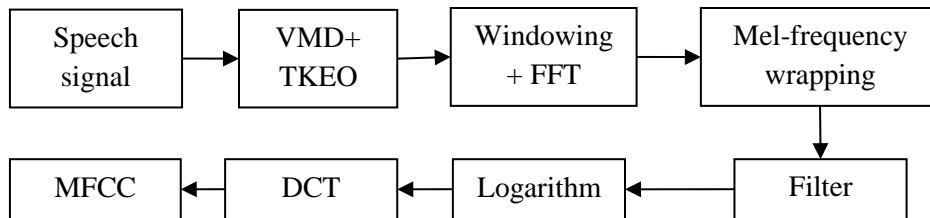


Fig. 3.4 Schema of MFCC Features Extraction for VMD-TKEO Method.

For the extraction of MFCC, the preprocessed signals $y(n)$ obtained from VMD-TKEO includes windowing (the duration of frame N_2 is 30ms and overlapping is equal to 20ms), calculating periodogram, mel frequency wrapping and lastly applying

discrete cosine transform [38]. The Fast Fourier Transform is computed using the DFT to convert the time domain signal into the frequency domain. The DFT of N_2 samples frame is defined as in Eqn. 3.6 [109],

$$Y_h = \sum_{n=0}^{N_2-1} y(n) e^{-\frac{j2\pi hn}{N_2}} \quad h = 0, 1, 2, 3 \dots \dots N_2 - 1. \quad (3.6)$$

Next is Mel-frequency wrapping stage which converts frequency (f is raw speech signal frequency) scale to Mel scale using Eqn. 3.7 [109],

$$M(f) = 1125 \ln\left(1 + \frac{f}{700}\right) \quad (3.7)$$

In MFCC most of the information is concentrated in the initial coefficients. So in this work first three mel frequency coefficients MFCC1, MFCC2 and MFCC3 are selected. Further, the statistical measures mean and variance of selected mel frequency coefficients are extracted as features.

3.1.3.4 Statistical Measure

Statistical measures are used to increase the computational efficiency of classifiers. These measures are computed from the pitch frequency and MFCC. In this work two statistical measures mean and variance are calculated. For a sequence r_i , where $i=1, 2, 3, \dots, m$, and m is sequence length. The mean (μ) and variance (σ^2) are calculated using Eqn. 3.8 and Eqn. 3.9 [112],

$$\mu = \frac{1}{m} \sum_{i=1}^m r_i \quad (3.8)$$

$$\sigma^2 = \frac{1}{m-1} \sum_{i=1}^m (r_i - \mu)^2. \quad (3.9)$$

3.1.4 Feature Selection

The classifier's accuracy can be optimized using the appropriate feature selection method. Generally, feature selection methods depend on the data type and data variability. In this work, the KW test is used for the feature selection. The probabilistic p-value of the KW test is used to examine features statistically [107]. In this work, the KW test is performed for the angry and neutral emotions class, and the selected feature set is commonly used for all the test sets. The selected features are energy, mean of pitch frequency (mPitch), variance of pitch frequency (vPitch), and mean and variance of MFCC1, MFCC2, and MFCC3. These features are chosen based on their ability to capture both spectral and temporal characteristics of speech signals. The MFCC coefficients are particularly effective for representing the shape of the vocal tract, which is crucial for emotion recognition. Energy and pitch features provide additional prosodic information, which enhances the ability to differentiate between emotions in binary classification tasks.

3.1.5 Classification

Supervised learning is employed in this study to train the SER model. During the training phase, the output is mapped to the input, and the learning parameters are adjusted to achieve high accuracy. There are various machine learning algorithms available that can be utilized with different types of data. Additionally, the selection of the optimal classifier for the dataset is crucial for achieving the highest accuracy possible for the SER model. In this work, the SVM classifier is used for training and testing. For the small size of the datasets, the SVM classifier proved more efficient. In this experiment, the SVM classifier variants Linear SVM (LSVM), Optimizable SVM (OSVM), Coarse Gaussian SVM (CGSVM), Quadratic SVM (QSVM), Cubic SVM (CSVM), Fine Gaussian SVM (FGSVM), and medium Gaussian SVM (MGSVM) are tested for selected feature set [53], [90].

3.2 Results and Discussion

This study utilizes the RVDESS dataset to investigate the effectiveness of the proposed approach for SER. The RVDESS dataset is well known for its high complexity and variability, as discussed in detail in chapter 1. To extract the meaningful information, the speech signals were decomposed into modes using the VMD and EMD technique. Subsequently, a nonlinear TKEO operator is applied to each of the modes, and resulting preprocessed signals are used for feature extraction.

Table 3.1 Total Features Extracted from Each Speech Signal.

| Name of Features Extracted | Number of Features Extracted from Each Speech Signal |
|-------------------------------------|---|
| Mean of MFCC1, MFCC2, and MFCC3 | 15 |
| Variance of MFCC1, MFCC2, and MFCC3 | 15 |
| mPitch | 5 |
| vPitch | 5 |
| Energy | 5 |

Table 3.2 Binary Class Emotions Accuracy Comparison for VMD and EMD.

| Emotions | Decomposition Method | SVM Classifier Variants Recognition Rate (%) | | | | | | |
|-------------------|---------------------------------|---|-------------|-------------|--------------|--------------|--------------|-------------|
| | | LSVM | QSVM | CSVM | FGSVM | MGSVM | CGSVM | OSVM |
| Angry, Neutral | VMD | 100 | 100 | 100 | 90 | 99.7 | 100 | 100 |
| | EMD | 71.5 | 72.2 | 71.2 | 66.7 | 71.5 | 69.4 | 71.9 |
| Angry, Happy | VMD | 93.8 | 97.1 | 95.6 | 84.2 | 95.6 | 90.4 | 95.3 |
| | EMD | 59.5 | 60.8 | 54.6 | 50.4 | 59.8 | 60.8 | 60.3 |
| Angry, Calm | VMD | 100 | 100 | 100 | 88.8 | 100 | 100 | 100 |
| | EMD | 62.8 | 63.4 | 61.0 | 51.3 | 62.8 | 58.9 | 63.4 |
| Happy, Sad | VMD | 92.5 | 93.1 | 91.5 | 85.4 | 92.1 | 90.6 | 92.2 |
| | EMD | 70.5 | 68.9 | 69.2 | 59.3 | 68.9 | 64.8 | 69.7 |

To determine the optimal number of modes, our experiment evaluated multiple modes and their corresponding features. As a result, proposed SER design employs five modes. The features used to design the SER system are described in section 3.1.3. The energy, pitch frequency, and MFCC, are extracted from each preprocessed signal. The energy and statistical variants of pitch frequency, MFCC1, MFCC2, and MFCC3, are used as features. The statistical significance of the extracted features is determined using the p -value obtained from the KW test.

Feature extracted from each speech signal is shown in Table 3.1. The features extracted from the preprocessed signal demonstrate a strong potential for accurately classifying emotions. This work tests variants of SVM classifier LSVM, QSVM, CSVM, FGSVM, MGSVM, CGSVM and OSVM for the selected feature set. The proposed model is evaluated for two-class, three-class, and four-class emotions test sets. The classification results are obtained using 10-fold cross-validation. In cross-validation, the data is randomly divided into q ($q=1, 2, 3, \dots$) subgroups; $q-1$ groups are used for training in each validation, and the left one is used for testing.

The Table 3.2 compares the performance of two methods based on VMD and EMD, for emotion recognition across different emotion pairs using various SVM classifier variants. For the Angry and Neutral combination, VMD based method achieves near-perfect recognition rates, with most classifiers reaching 100%, except FGSVM, which achieves 90%. In contrast, EMD based method performs significantly worse, with recognition rates ranging between 66.7% and 72.2%, highlighting its limitations in extracting features that effectively distinguish between these emotions. Similarly, for the Angry and Happy pair, VMD based method demonstrates superior performance, with recognition rates ranging from 84.2% to 97.1%, whereas EMD based method struggles with values between 50.4% and 60.8%. This indicates that VMD based method is far more robust in handling this emotion pair compared to EMD based method.

For the Angry and Calm combination, VMD based method again delivers exceptional results, achieving 100% accuracy for most classifiers except FGSVM, which achieves 88.8%. On the other hand, EMD based method achieves much lower recognition rates, ranging from 51.3% to 63.4%, further emphasizing its inability to effectively separate features for these emotions. In the case of Happy and Sad, VMD based method continues to show strong performance, with recognition rates between 85.4% and 93.1%, demonstrating its effectiveness even for subtle emotional differences. EMD based method, however, achieves only moderate recognition rates, ranging from 59.3% to 70.5%, which indicates its struggles in handling closely related emotions.

Overall, VMD based method consistently outperforms EMD based method across all emotion pairs and classifier variants. VMD's superior performance can be attributed to its ability to provide better signal decomposition, resulting in highly discriminative features for the SVM classifiers. In contrast, EMD appears to suffer from limitations such as mode-mixing, which reduces its effectiveness in emotion recognition tasks. Among the classifiers, LSVM and QSVM generally perform better, particularly with VMD, while FGSVM consistently delivers lower accuracy across both methods. These results clearly demonstrate that VMD is the more effective decomposition method for speech emotion recognition, enabling significantly higher recognition rates compared to EMD.

Table 3.3 Proposed Model Accuracy for Three Class Emotions.

| Emotions | SVM Classifier Variants Recognition Rate (%) | | | | | | |
|-----------------------------|---|-------------|-------------|--------------|--------------|--------------|-------------|
| | LSVM | QSVM | CSVM | FGSVM | MGSVM | CGSVM | OSVM |
| Angry, Calm, Neutral | 89.4 | 86.2 | 85.8 | 71.0 | 87.9 | 80.0 | 86.9 |
| Angry, Happy, Neutral | 83.4 | 86.5 | 85.2 | 69.4 | 85.7 | 81.5 | 85.7 |
| Angry, Happy, Calm | 88.2 | 92.2 | 92.1 | 77.2 | 91 | 84.3 | 91.1 |

Table 3.4 Proposed Model Accuracy for Four Class Emotions.

| Emotions | SVM Classifier Variants Recognition Rate (%) | | | | | | |
|--------------------------------------|---|-------------|-------------|--------------|--------------|--------------|-------------|
| | LSVM | QSVM | CSVM | FGSVM | MGSVM | CGSVM | OSVM |
| Angry, Calm, Neutral, Happy | 78.7 | 81.7 | 80.6 | 66.3 | 79.4 | 72.6 | 79.6 |

Table 3.5 Emotion Recognition Accuracy for the Different Test Sets.

| Emotion Class | Accuracy for Different Emotions (%) | | | |
|-----------------------------|-------------------------------------|-------|---------|-------|
| | Angry | Happy | Neutral | Calm |
| Angry, Neutral | 100 | - | 100 | - |
| Angry, Happy | 96.39 | 97.9 | - | - |
| Angry, Calm | 100 | - | - | 100 |
| Angry, Calm, Neutral | 100 | - | 70.83 | 87.89 |
| Angry, Happy, Neutral | 96.39 | 83.24 | 72.91 | - |
| Angry, Happy, Calm | 96.9 | 86.38 | - | 93.15 |
| Angry, Calm, Neutral, Happy | 96.39 | 83.33 | 45.83 | 85.26 |

Table 3.6 Emotion Recognition Results from Raw Speech Signal and VMD-TKEO Preprocessed Speech Signal.

| Emotion Classes | Accuracy with QSVM classifier (%) | |
|-----------------------------|-----------------------------------|-------------------------------------|
| | Raw Speech Signal | VMD-TKEO Preprocessed Speech Signal |
| Angry, Neutral | 76.7 | 100 |
| Angry, Happy | 63.4 | 97.1 |
| Angry, Calm | 83.2 | 100 |
| Angry, Calm, Neutral | 61.5 | 86.2 |
| Angry, Happy, Neutral | 54.5 | 86.5 |
| Angry, Happy, Calm | 63.2 | 92.2 |
| Angry, Calm, Neutral, Happy | 47.2 | 81.7 |

As the number of emotion classes increases in Table 3.3 and Table 3.4, the recognition rate of the proposed SER model is compromised. In Table 3.3, the test set having angry, happy, and calm emotions shows the highest accuracy of 92.3%, while the test set having natural emotion shows comparatively low accuracy of 89% and 85.9%. Table 3.2 and Table 3.3 show the type of emotion also affects the recognition rate of the SER model. Table 3.3 and Table 3.4 show that the multiclass test set with a neutral

emotion class achieves minimum accuracy. Finally, it can be concluded that the accuracy of the SER model is highly correlated with the type of emotion, intensity, and the number of emotion classes of emotion.

For the quantitative analysis of the proposed model, the best emotion accuracy of each test set is included in Table 3.5. Table 3.5 shows the angry emotion class's highest recognition accuracy for all the test sets. The selected features outperform for classifying high-energy and low-energy emotions, whereas they have the lowest accuracy for normal-energy emotions. The proposed architecture improves the recognition accuracy for all classes except the neutral emotion class. The used dataset recorded in high and low intensity, so normal energy emotions in high intensity behave like high energy emotions and in low-intensity act like low energy emotions.

Table 3.6 presents the emotion recognition experiment results without (raw speech signal) and with the proposed method (VMD-TKEO). In both experiments, extracted features are tested on variants of the SVM classifier. The QSVM classifier achieved the highest accuracy across all combinations in both cases. Table 3.6 shows that the VMD-TKEO based approach provided better performance for recognizing all speech emotions than raw speech signals.

Table 3.7 Comparison Table of Proposed Method with Existing Methods.

| Authors | Number of Emotion Classes | Method | Accuracy (%) | Dataset |
|------------------------|----------------------------------|---|---------------------|----------------|
| R. Jannat et al [72] | 2 | CNN based architecture | 66.41 | RAVDESS |
| M. A. Jalal et al [74] | 2 | BLSTM and CNN based capsule network | 79.5 | RAVDESS |
| B. Zhang et al [89] | 4 | Multi-task learning approach with six binary claassifiers | 54.76 | RAVDESS |
| Proposed Method | 2 | VMD+TKEO+QSVM | 100 | RAVDESS |
| Proposed Method | 3 | VMD+TKEO+QSVM | 92.2 | RAVDESS |
| Proposed Method | 4 | VMD+TKEO+QSVM | 81.7 | RAVDESS |

Table 3.6 presents the emotion recognition experiment results without (raw speech signal) and with the proposed method (VMD-TKEO). In both experiments, extracted features are tested on variants of the SVM classifier. The QSVM classifier achieved the highest accuracy across all combinations in both cases. Table 3.6 shows that the VMD-TKEO based approach provided better performance for recognizing all speech emotions than raw speech signals.

Table 3.7 presents a comparison between the proposed SER model and existing models. R. Jannat and M.A. Jalal suggested the solutions with an accuracy of 66.41% and 79.5% for the binary class classification. R. Jannat et al. [72] used a convolutional neural network, while M.A. Jalal et al [74] proposed a temporal modelling framework. B. Zhang et al. [89] proposed the multi-task learning approach with six binary classifiers for the four-class classification, gaining 54.76% accuracy.

The proposed method VMD-TKEO exhibits superior classification performance for two, three, and four-class emotions classification with 100%, 92.2%, and 81.7% accuracy, respectively. This performance is obtained due to the ability of VMD-TKEO to analyze signals with non-uniform frequency content, where traditional signal processing techniques find limitations. As far as the proposed method limitation concern the VMD-TKEO is computation expensive, especially for large data sets. This limitation can make it challenging to use in real-time applications or on low-power devices.

3.3 Summary

The SER architecture based on VMD outperformed the EMD-based approach, achieving a higher accuracy of 100% for binary classification. In this architecture, VMD with the TKEO energy operator is explored to analyze speech signal. Features are extracted from the VMD-TKEO processed signal and statistically examined using the p -value of the KW test. The selected feature set is evaluated over variants of the SVM classifier for analyzing two-class, three-class, and four-class emotions categories. The QSVM provides the best results for all emotion recognition categories. The highest achieved recognition rates for two-class, three-class, and four-class categories are 100%, 92.2%, and 81.7%, respectively. This performance is also better as compared to other existing works. It is observed that the performance of the VMD-based approach degrades as the number of emotion classes increases. This might be due to the presence of less informative modes.

CHAPTER 4

MULTICLASS SPEECH EMOTION RECOGNITION

In previous chapter, VMD was combined with TKEO, achieving the highest accuracy for binary classification. However, as the number of emotion classes increased, the model's accuracy decreased. Therefore, in this chapter, a multiclass speech emotion recognition approach is presented, utilizing VMD with adaptive mode selection based on energy information. Instead of directly analysing the raw speech signal, this work emphasizes the pre-processing stage to enhance emotion recognition performance. The VMD effectively separates the signal into multiple modes, each representing distinct frequency components. The energy of each mode is then calculated to identify the dominant modes that contribute most significantly to the speech signal's characteristics. These dominant modes are subsequently used for signal reconstruction, ensuring that the processed signal retains its most relevant emotional features. After reconstruction, the signal is divided into frames, allowing for the extraction of both prosodic and spectral features. Following feature extraction, the ReliefF algorithm is applied for feature selection. The selected feature set is then used to train a fine-tuned KNN classifier for emotion identification. The proposed framework demonstrates robust performance across various datasets, achieving accuracies of 93.8%, 95.8%, 93.4%, and 83.10% on RAVDESS-speech, Emo-DB, EMOVO, and IEMOCAP datasets, respectively. These results highlight the framework's effectiveness in handling multilingual and diverse speech emotion recognition tasks. Furthermore, the proposed method has also proven to be robust for three languages: English, German, and Italian, with language sensitivity as low as 2.4% compared to existing methods. The removal of noise using dominant mode energy based method significantly improves the sensitivity of proposed model. This technique effectively suppresses unwanted noise components. As a result, the speech signals become much cleaner and more distinct. Cleaner signals enhance the model's ability to detect subtle emotional cues. Overall, the sensitivity of the proposed SER model improves significantly. Combinig VMD with dominant mode energy significantly improves the predictability of speech signal.

4.1 Multiclass SER Model

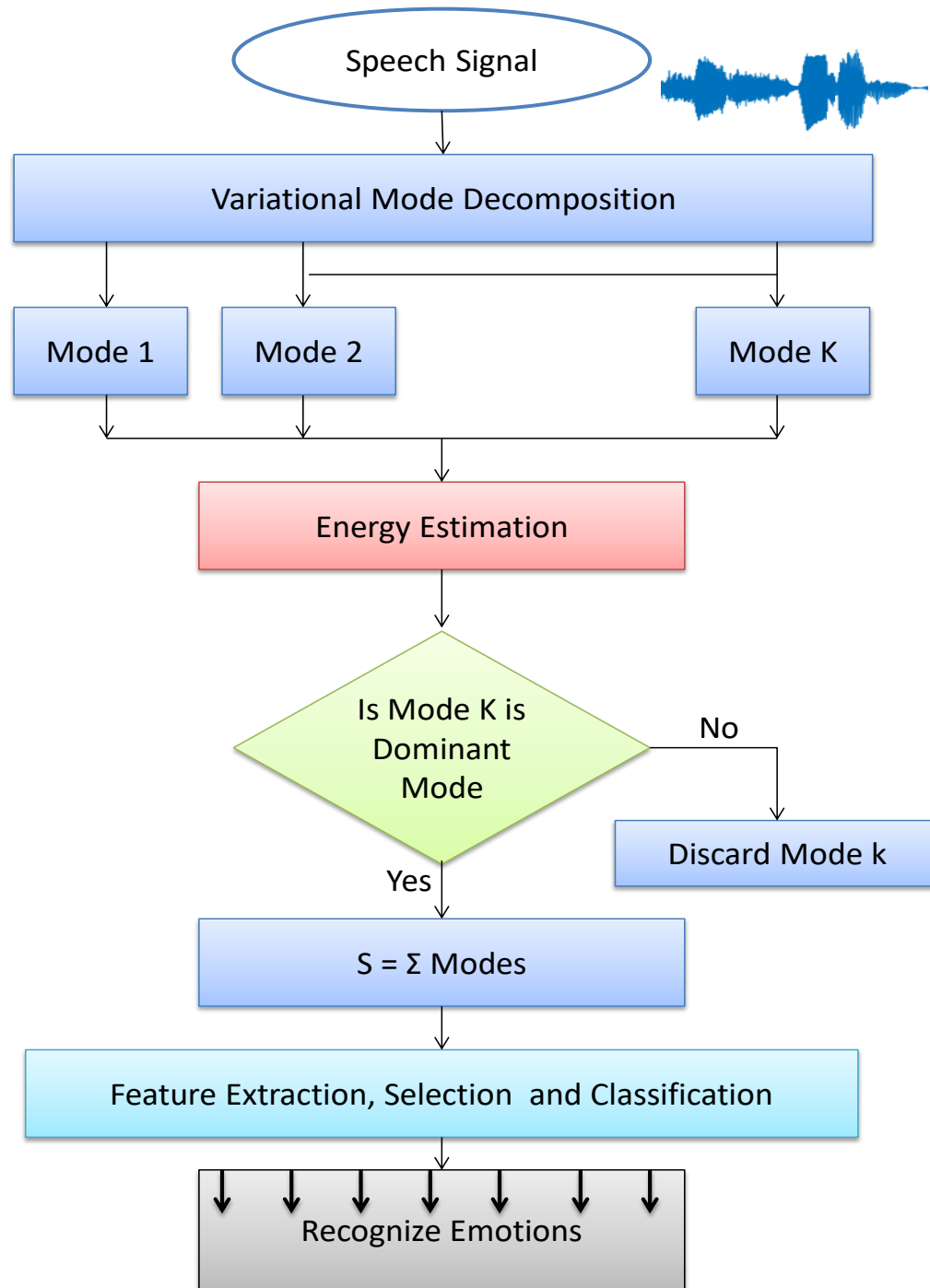


Fig. 4.1 Block Diagram of Proposed Framework. Where, K is Number of Modes.

The structure of the proposed framework is illustrated in Fig. 4.1. The complete framework is divided in three key steps (i) signal preprocessing, (ii) prosodic and spectral feature computation and selection (iii) classification. The comprehensive description of these three steps is outlined in the subsequent sections.

The input raw speech signal decomposition, energy estimation and signal reconstruction is presented in algorithm 1.

Algorithm 1 Algorithm for dominant Mode selection and signal reconstruction.

Input: Raw audio signal f

Output: Reconstructed signal $s[n]$ using the 6 dominant energy modes for each audio signal

1: Initialized desired frequency range for reconstructed signal, $s[n]=0$

2: **for** $i=1: K$ **do**

3: $E_i = \text{energy}(\text{Mode}_i)$

4: **If** E_i is dominant **then**

5: Mode_i is reconstructed signal component

6: $s[n] = s[n] + \text{Mode}_i$

7: **else**

8: Mode_i is undesired signal frequency component

9: **end if**

10: **end for**

11: Get $s[n]$

4.1.1 Signal Preprocessing

Signal preprocessing involves removing unwanted parts from the input raw speech signal to make the signal more predictable. In this work, preprocessing includes decomposition, energy estimation, and reconstruction of the signal. A detailed description of each step of preprocessing is available in the following sections.

During the initial preprocessing step, the VMD is employed on the raw speech signal to compute its modes. VMD decomposes the real valued signal in number of modes with certain specific properties. This decomposition method is inherently non recursive by nature and every mode is centralized on the central frequency. In proposed framework

each raw speech signal is decomposed into nine modes using VMD. For example, the all nine mode of input raw speech signal are presented in Fig. 4.2.

The VMD decomposes the input raw speech signal into nine modes. Following decomposition, the energy of each mode is calculated using Eqn. 4.1 [108],

$$E = \sum_{n=1}^N |u_k(n)|^2 \quad (4.1)$$

Where, N is input raw speech signal length and n is discrete point.

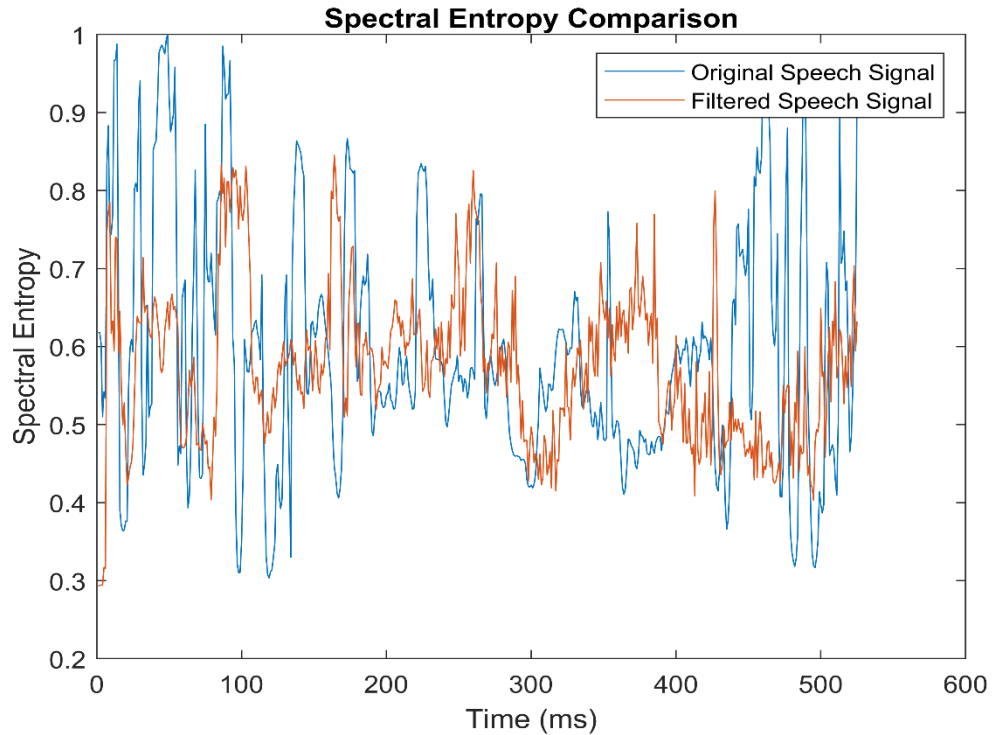


Fig. 4.2 The Spectral Entropy Comparison of the Raw Speech Signal and the Reconstructed Signal.

For example, the mode energy of angry and sad emotion from three datasets RAVDESS, EMOVO, and Emo-DB are presented in Table 4.1. In the angry emotion of the RAVDESS dataset, modes 3, 5, 6, 7, 8, and 9 have higher energy compared to the mode 1, 2, and 4. Similarly, in the sad emotion of the RAVDESS dataset modes 4 to 9 have higher energy. The higher energy modes are considered as the dominant mode and

are used for the signal reconstruction. In the EMOVO dataset modes 3,5,6,7,8,9 for angry emotion and mode 2,3,4,7,8,9 for sad emotion are used as the dominant mode. Further, in the Emo-DB dataset modes, 3, 5, 6, 7, 8, 9 for angry emotion and modes 2, 3, 6, 7, 8, 9 for sad emotion are used as the dominant mode. In the IEMOCAP dataset modes 3,4,6,7,8,9 for angry emotion and mode 4 to 9 for sad emotion are used as the dominant mode.

Table 4.1 For Example, Energy of Angry and Sad Emotion from four Datasets RAVDESS, EMOVO, Emo-DB, and IEMOCAP. E1 to E9 Represents the Energy of Modes from One to Nine.

| Dataset | Emotion | E1 | E2 | E3 | E4 | E5 | E6 | E7 | E8 | E9 |
|----------------|----------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|
| RAVDESS | Angry | 0.10 | 0.16 | 0.41 | 0.33 | 0.87 | 1.15 | 4.44 | 12.68 | 13.28 |
| | Sad | 0.01 | 0.01 | 0.01 | 0.03 | 0.02 | 0.04 | 0.08 | 0.77 | 1.11 |
| EMOVO | Angry | 1.63 | 1.44 | 1.45 | 1.22 | 3.56 | 11.22 | 29.37 | 55.10 | 176.24 |
| | Sad | 0.71 | 1.99 | 1.24 | 1.65 | 0.93 | 1.06 | 3.73 | 27.44 | 17.89 |
| Emo-DB | Angry | 6.85 | 12.31 | 25.12 | 23.69 | 46.72 | 56.20 | 46.90 | 109.6 | 43.87 |
| | Sad | 5.38 | 6.81 | 8.66 | 3.67 | 2.97 | 4.31 | 39.88 | 106.20 | 167.87 |
| IEMOCAP | Angry | 0.48 | 1.37 | 3.19 | 3.12 | 2.55 | 11.26 | 29.80 | 60.08 | 45.82 |
| | Sad | 0.01 | 0.02 | 0.04 | 0.19 | 0.07 | 0.71 | 0.67 | 1.91 | 8.20 |

The calculated energy is utilized for the mode selection. In this work six dominant modes of each raw speech signal having high energy level are selected for the signal reconstruction. The spectral entropy in Fig. 4.3 illustrates that reconstructed speech signal have lower entropy compared to the raw speech signal. Since the voiced part is concentrated in a higher energy band, the reconstructed signal is comparatively more predictable. The signal s is reconstructed using Eqn. 4.2,

$$s = \sum_{k=i} u_k(n) \quad (4.2)$$

Where i is dominant mode number and can vary from one to nine.

The root mean square (RMS) value and spectral entropy are calculated to assess the predictability of reconstructed speech signal. For example, the angry mode signal from the RAVDESS dataset is used to showcase the predictability of the reconstructed signal. The obtained RMS value of raw speech signal and reconstructed speech signal is 0.012 and 0.082 respectively. The higher RMS value indicates that the reconstructed signal is more structured and less noisy because noise tends to lower the RMS and spread out the energy, making the signal less predictable. The spectral entropy comparison is presented in Fig. 4.3. In Fig. 4.3, the blue graph representing the raw speech signal shows higher spectral entropy, while the other colored graph representing the reconstructed signal shows lower spectral entropy. This indicates that the preprocessing process has reduced randomness or noise, making the signal more predictable. Since the voiced part is concentrated in a higher energy band, the reconstructed signal is comparatively more predictable.

4.1.2 Feature Extraction, Selection and Classification

The reconstructed signal is segmented into frames. In this experiment, 80 millisecond frame with 75 percent overlapping is used for the feature extraction. After segmentation hamming window is used for smoothing the edges [103]. The speech features are computed from each frame. The extracted features and their short description are as follow.

Spectral skewness is calculated to capture the asymmetry of the spectral distribution of a signal, indicating whether the spectral energy is concentrated more towards the higher or lower frequencies. Spectral spread is computed to measures the extent of dispersion of spectral components, indicating how widely spread out the frequencies are within the spectrum. Harmonic ratio quantifies the presence of harmonics

in a signal, often used to distinguish between harmonic and non-harmonic sounds. Another feature Gammatone cepstral coefficients (GTCC) are also calculated to assess the auditory system's response to sound. They are obtained by passing the audio signal through a bank of gammatone filters, which mimic the frequency response of the human hearing system. The output obtained from filters is processed using cepstral analysis techniques to get the features. GTCCs offer a concise portrayal of the spectral features of the audio signal, encapsulating crucial details regarding its timbre and frequency composition. The first derivative of GTCC is also used as feature [103]. MFCC acquire the frequency content of the audio signal in a way that is perceptually meaningful, emphasizing important features while suppressing irrelevant variations. First and second derivative of MFCC is also calculated [109]. Pitch frequency estimation algorithms analyze the temporal variations in the waveform or the spectrum of the signal to identify periodic patterns corresponding to the fundamental frequency [4]. Mel spectrum, spectral crest, spectral entropy, spectral kurtosis, and spectral centroid are also evaluated from each frame to improve the recognition accuracy [5]. To reduce classifier computation complexity and enhance SER accuracy these computed features undergo a feature selection process to acquire an optimal feature set.

To attain the optimal feature set, a rank-based feature selection approach is implemented. In this framework ReliefF algorithm is employed for the feature optimization. This algorithm randomly selects instances from both within a particular category and across various categories. During the process of selecting R data instances out of the total n instances, this algorithm computes the feature score f_r for every input feature ($r=1,2,3,\dots,d$), with d representing the total number of input features. The f_r score is used for the dominant feature selection. This computation follows the formula outlined in Eqn. 4.3 of the specified reference [110],

$$f_r = \frac{1}{c} \sum_{q=1}^R \left(-\frac{1}{m_q} \sum_{x_l \in NS(q)} d(X(q, r) - X(l, r)) + \sum_{z \neq q} \frac{1}{h_{qz}} \frac{p(z)}{1-p(z)} \sum_{x_l \in NH(q, z)} d(X(q, r) - X(l, r)) \right) \quad (4.3)$$

Here $NS(q)$ and $NH(q, y)$ are defined as the nearest instances of x_q in identical category of size m_q and in category z of size h_{qz} respectively. The c is number of category and $p(y)$ represents the ratio of instances for z category.

The selected feature is used to train the classifier. In this experiment KNN classifier based on Euclidian distance is tested for emotion classification. The Euclidian distance is calculated between predefined class and each varying sample. In the implemented framework, fine KNN are utilized for emotion identification, utilizing a

single sample to differentiate the data. The Euclidian distance is computed using Eqn. 4.4 [88],

$$D = \sqrt{\sum_{e=1}^R (y_{1e} - y_{2e})^2} \quad (4.4)$$

Where e represents the discrete points in each sample, and y_1, y_2 are input samples. To evaluate the effectiveness of the implemented framework, precision, recall, and $F1$ -score are calculated for each class and experiment.

4.2 Results and Discussion

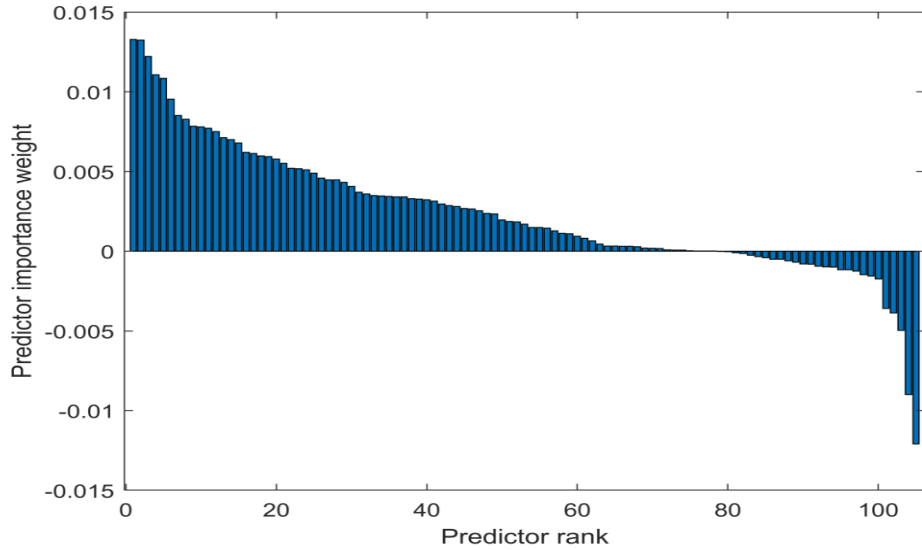


Fig. 4.3 Predictor Rank Bar of RAVDESS Experiment.

In this work four experiments using publically available four speech emotion database RAVDESS, EMOVO, Emo-DB, and IEMOCAP are performed. For the experiment, three different language datasets are chosen to test the robustness of implemented method. In each experiment raw speech signal undergo the preprocessing, feature extraction and classification stage.

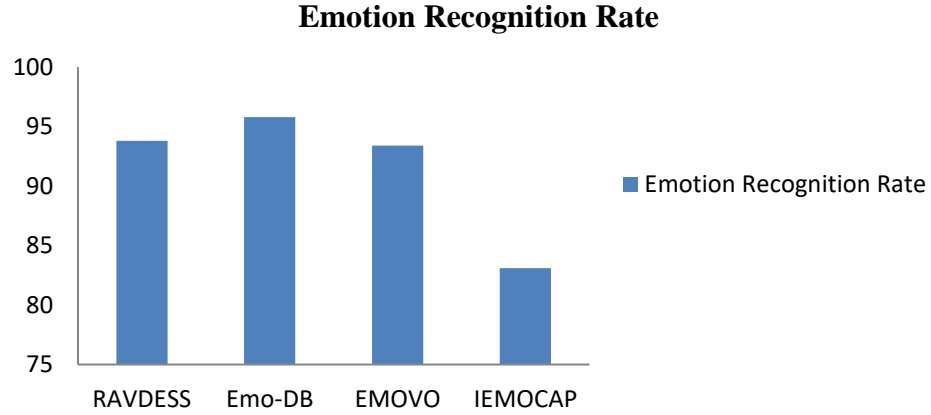


Fig. 4.4 Recognition Rate of RAVDESS, Emo-DB and EMOVO Experiment.

For the preprocessing stage, the raw speech signal undergoes decomposition into nine modes using VMD, as detailed in the preceding section. After decomposition the energy of each mode is calculated. Modes having higher energy considered as the dominant mode and used for the signal reconstruction. In this work six dominant mode are considered for the signal reconstruction. The reconstructed signal is utilized for subsequent processing. The prosodic and spectral features are extracted from each reconstructed signal. For the feature selection rank based ReliefF algorithm is used. In this work, 50 features with high predictor ranks and positive weights are selected to train the classifier, except in the IEMOCAP experiment. In the IEMOCAP experiment, only 36 features with positive weights are obtained, and all of them are selected for training and testing the classifier. For example, the predictor rank bar of RAVDESS experiment is presented in Fig. 4.4. The KNN classifier is used for testing the selected feature sets of each experiment. In this work, the classifier is trained using 10-fold cross-validation. The bar graph illustrating the accuracy of all experiments is depicted in Fig. 4.5.

The results highlight the effectiveness of the selected feature sets, showing that the dominant modes contribute significantly to improving classification performance. In addition, the classifier demonstrates consistent results across various datasets, emphasizing its adaptability. Future work may explore the application of different classifiers, such as support vector machines or deep learning models, to further improve accuracy. Furthermore, the influence of various preprocessing techniques on the feature extraction process could be explored to enhance the robustness of the system in noisy environments. Overall, the proposed methodology offers a promising framework for emotion recognition in speech.

| | | | | | | | | |
|---|-------|-------|-------|-------|-------|-------|-------|-------|
| 1 | 34496 | 187 | 294 | 324 | 377 | 155 | 294 | 367 |
| 2 | 140 | 33953 | 302 | 175 | 194 | 361 | 411 | 227 |
| 3 | 251 | 307 | 35068 | 296 | 331 | 161 | 374 | 373 |
| 4 | 270 | 201 | 311 | 31682 | 295 | 173 | 345 | 361 |
| 5 | 351 | 302 | 290 | 318 | 31940 | 201 | 372 | 479 |
| 6 | 112 | 384 | 149 | 114 | 145 | 15119 | 257 | 195 |
| 7 | 227 | 433 | 336 | 318 | 310 | 305 | 32515 | 348 |
| 8 | 323 | 269 | 402 | 377 | 453 | 237 | 381 | 30355 |
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |

True Class

Predicted Class

Fig. 4.5 Confusion Matrix of RAVDESS Experiment.

| | | | | | | | |
|---|-------|-------|-------|-------|-------|-------|-------|
| 1 | 10945 | 68 | 218 | 143 | 261 | 146 | 158 |
| 2 | 45 | 13511 | 40 | 93 | 106 | 146 | 151 |
| 3 | 242 | 60 | 11236 | 128 | 145 | 123 | 99 |
| 4 | 146 | 133 | 100 | 10975 | 115 | 120 | 147 |
| 5 | 219 | 141 | 126 | 84 | 11887 | 251 | 158 |
| 6 | 131 | 186 | 117 | 106 | 228 | 11840 | 141 |
| 7 | 146 | 213 | 95 | 170 | 150 | 169 | 13357 |
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 |

True Class

Predicted Class

Fig. 4.6 Confusion Matrix of EMOVO Experiment.

| | | | | | | | |
|---|------|------|------|------|------|------|------|
| 1 | 4960 | 13 | 15 | 34 | 109 | 2 | 16 |
| 2 | 3 | 3308 | 7 | 20 | 13 | 17 | 101 |
| 3 | 26 | 21 | 2307 | 11 | 10 | 13 | 18 |
| 4 | 18 | 15 | 10 | 2240 | 22 | 14 | 11 |
| 5 | 88 | 14 | 12 | 42 | 2578 | 5 | 24 |
| 6 | 1 | 22 | 11 | 8 | 2 | 3896 | 29 |
| 7 | 16 | 79 | 8 | 16 | 14 | 24 | 2673 |
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 |

True Class

Predicted Class

Fig. 4.7 Confusion Matrix of Emo-DB Experiment.

| | | | | |
|---|--------|--------|--------|--------|
| 1 | 168365 | 6011 | 15970 | 9654 |
| 2 | 5043 | 187577 | 8672 | 8708 |
| 3 | 15136 | 7048 | 144972 | 20844 |
| 4 | 7945 | 7183 | 21823 | 157049 |
| | 1 | 2 | 3 | 4 |

True Class

Predicted Class

Fig. 4.8 Confusion Matrix of IEMOCAP Experiment.

Table 4.2 Precision, Recall and F1 Score for the RAVDESS Experiment.

| Emotions | Precision (%) | Recall (%) | F1 score (%) |
|-----------------|----------------------|-------------------|---------------------|
| Calm | 94.14 | 94.93 | 94.53 |
| Angry | 95.37 | 94.52 | 94.94 |
| Sad | 93.03 | 93.45 | 93.24 |
| Disgust | 94.39 | 94.36 | 94.37 |
| Fearful | 94.28 | 94.18 | 94.23 |
| Surprised | 92.81 | 92.55 | 92.68 |
| Happy | 93.81 | 93.24 | 93.52 |
| Neutral | 90.46 | 91.76 | 91.11 |

Table 4.3 Precision, Recall and F1 Score for the EMOVO Experiment.

| Emotions | Precision (%) | Recall (%) | F1 score (%) |
|-----------------|----------------------|-------------------|---------------------|
| Joy | 92.17 | 91.67 | 91.92 |
| Disgust | 93.99 | 93.40 | 93.69 |
| Fear | 94.20 | 92.87 | 93.53 |
| Surprised | 92.20 | 94.04 | 93.11 |
| Neutral | 93.81 | 93.51 | 93.66 |
| Sad | 94.40 | 95.87 | 95.13 |
| Angry | 94.16 | 93.37 | 93.76 |

Table 4.4 Precision, Recall and F1 Score for the EMO-DB Experiment.

| Emotions | Precision (%) | Recall (%) | F1 score (%) |
|-----------------|----------------------|-------------------|---------------------|
| Boredom | 95.27 | 95.35 | 95.31 |
| Disgust | 97.34 | 95.88 | 96.60 |
| Fear | 94.47 | 96.13 | 95.29 |
| Happy | 93.81 | 93.30 | 93.55 |
| Neutral | 93.07 | 94.45 | 93.75 |
| Sad | 98.11 | 98.16 | 98.13 |
| Angry | 97.02 | 96.32 | 96.67 |

Table 4.5 Precision, Recall and F1 Score for the IEMOCAP Experiment.

| Emotions | Precision (%) | Recall (%) | F1 score (%) |
|----------|---------------|------------|--------------|
| Angry | 86.10 | 84.09 | 85.09 |
| Happy | 90.21 | 89.07 | 90.07 |
| Neutral | 76.12 | 77.03 | 76.12 |
| Sad | 80.11 | 81.10 | 80.11 |

Table 4.6 Recognition Rates of Various SER State of the Art Methods for Comparative Performance Analysis.

| Author | Database | Classifier | Accuracy (%) | Language Sensitivity (%) |
|---------------------------|--|-----------------------|----------------------------------|--------------------------|
| M. Sajjad, et al. [113] | EMO-DB RAVDESS Speech | RBFN CNN BILSTM | 85.57 77.02 | 8.55 |
| T. Özseven, et al. [96] | EMO-DB SAVEE EMOVO | SVM | 84.62 74.39 60.4 | 24.22 |
| A. Milton, et al. [98] | EMO-DB RAVDESS Speech EMOVO | MSVM | 81.5 64.31 73.3 | 8.20 |
| Soonil, et al. [99] | EMO-DB SAVEE RAVDESS Speech | Two-Stream CNN | 95 82 85 | 13 |
| Sengul, et al. [71] | EMOVO RAVDESS Speech EMO-DB SAVEE | SVM | 90.09 84.79 79.08 87.43 | 11.01 |
| Leila, et al. [93] | EMOVO EMO-DB | RNN SVM | 91.16 86.22 | 4.94 |
| Proposed framework | RAVDESS Speech EMO-DB EMOVO | Fine KNN | 93.80 95.80 93.40 | 2.40 |

The proposed framework achieves 93.8% SER accuracy for RAVDESS dataset. The confusion matrices obtained after training the classifier using all four datasets are presented in Fig. 4.6 to Fig. 4.9 (Emotions serial number is correspond to respective Precision, Recall and F1 Score table). Table 4.2 shows the precision, recall and F1 score of each emotions of RAVDESS experiment. According to RAVDESS experiment highest precision, recall and F1 score is obtained for the angry emotion. While, the lowest precision, recall and F1 score is obtained for the neutral category. Neutral emotion achieves lowest accuracy because available training samples are half compared to other emotion. It shows that the size of dataset affects the accuracy of classifier. Table 4.3 and Table 4.4 presents the precision, recall and F1 score of EMOVO and Emo-DB experiment. EMOVO experiment attains 93.4% emotion identification accuracy using KNN classifier. The highest recall result of 95.87%, 94.04%, and 93.51% were obtained for sad, surprised, and neutral respectively. The lowest precision result of 92.17%, 92.20% and 93.81% were obtained for joy, surprised, and neutral respectively. Emo-DB experiment attains highest 95.8% emotion identification accuracy using KNN classifier. The highest recall result of 98.16%, 96.32%, and 96.13% were obtained for sad, angry, and fear respectively. The lowest precision result of 93.07%, 93.81% and 94.47% were obtained for neutral, happy, and fear respectively. The proposed method is also proved efficient for the IEMOCAP dataset and provide 83.10% accuracy. Table 4.5 shows the precision, recall and f1 score for the IEMOCAP experiment.

Table 4.6 highlights the strengths and limitations of the proposed SER framework. It demonstrates high sensitivity for detecting happy emotion but struggles with neutral emotion, likely due to its subtle acoustic variations. The framework is most sensitive to the German language, reflecting robust feature alignment, and least sensitive to Italian, where linguistic or acoustic nuances might reduce performance. The accuracy disparity of just 2.4% between German, English and Italian datasets underscores the framework's robustness and adaptability across languages. This small variation suggests that while some linguistic dependencies exist, the model maintains a strong generalization capability for multilingual emotion recognition tasks.

The efficiency of the proposed method compared to the existing state of the art is depicted in Table 4.7. The machine learning and deep learning based recent SER framework is included to show the potential of proposed framework. The presented SER framework provide higher accuracy compared to [113], [98], [99], and [71] for RAVDESS speech database. The EMOVO experiment proved more efficient compared to [96], [98], and [93]. Further, Emo-Db experiment also gained higher accuracy compared to [113], [96], [98], [99], [71], [93]. Table 4.6, demonstrate that the accuracy of the SER framework varies significantly depending on the language. The framework presented by [93] shows the lowest language sensitivity 4.94%, while [96] shows the highest language sensitivity 24.22%. The performance evaluation demonstrates that the presented framework achieves the highest accuracy across all three datasets, exhibiting minimal language dependency

compared to existing state-of-the-art methods. Table 4.7 shows the comparison of the proposed method with the existing state of the art for the IEMOCAP dataset. The performance evaluation demonstrates that the presented framework achieves the highest accuracy across all four datasets and exhibits minimal language dependency for the acted dataset compared to existing state-of-the-art methods. The major advantage of the proposed method is that it renders the signal more predictable by removing the low-energy frequency band. For classification purposes, fine KNN is utilized, requiring fewer parameters for training. This, in turn, makes the proposed method lightweight.

Table 4.7 Recognition Rates of Various SER State of the Art Methods for Elicited Dataset IEMOCAP for Comparative Performance Analysis.

| Author | Database | Classifier | Method | Accuracy (%) |
|---------------------------|----------------|--|--|--------------|
| XU et al. [114] | IEMOCAP | Attention-based CNN | Head Fusion based on the multi-head attention mechanism is used to improve the accuracy of SER | 76.36 |
| Issa et al. [115] | IEMOCAP | One-dimensional CNN | Extracts MFCC, chromagram, mel-scale spectrogram, Tonnetz representation, and spectral contrast features from sound files and uses them as inputs for the one-dimensional CNN. | 64.30 |
| Liu, et al. [116] | IEMOCAP | CNN and attention-based bidirectional long short-term memory network | Time-domain filter and frequency-domain filter are used to process the triple-channel log-Mel spectrograms respectively in order to increase the diversity of features | 70.27 |
| Proposed framework | IEMOCAP | Fine KNN | VMD decomposition, and energy estimation | 83.10 |

4.3 Summary

The proposed method shows that the VMD decomposition and dominant mode selection approaches make the signal predictable. The results obtained from different experiments prove that most of the information exists in dominant modes. While the lower energy modes contain undesired information. The feature extracted from the preprocessed signal is more closely associated with emotion rather than speech content. The feature optimization algorithm helps to select the potential features for emotion classification. The selected feature set are used to train and test the KNN classifier. To check the robustness, the implemented method is tested on four dataset RAVDESS, EmoDB, EMOVO, and IEMOCAP. The proposed framework proves robust for three languages: English, German, and Italian. The result of all three experiments is compared with the existing state of art. The proposed method concentrates on removing the less informative modes, but noise can mask the subtle changes in emotions.

CHAPTER 5

MULTILINGUAL SPEECH EMOTION RECOGNITION

Emotions are conveyed through subtle changes in speech parameters, such as pitch, loudness, and speaking rate. Noise in the speech signal can mask these changes, making it challenging to recognize emotions accurately. Denoising can help remove the noise and reveal subtle changes in the speech signal. The MLSER system can produce more consistent results by eliminating the noise, regardless of the recording conditions. Hence, this work focuses on the preprocessing of input speech data. This study proposes a novel decomposition-based architecture for MLSER. The architecture includes silence removal, mode tuning, signal reconstruction, feature extraction, feature optimisation and classification. In preprocessing, the silence part is removed using short-time energy and spectral centroid. After that, VMD is applied for signal decomposition, where the improved Bhattacharyya distance is explored for the decomposition mode tuning. The tuned modes are examined for noise removal, and the signal is reconstructed using denoised modes. The spectral and prosodic features are computed from the reconstructed signal. The optimized features are obtained from the extracted features using the ReliefF algorithm. Finally, the fine k-nearest neighbor classifier is explored with optimized features to identify the emotions. For the experiment, three publicly available emotion databases, namely the English language-based RAVDESS, German language-based EMO-DB and Italian emotional speech database EMOVO, are used. The proposed method yielded 90.7%, 94% and 91.1% accuracy for English, German, and Italian language-based database, respectively. A multilingual database is created with these three databases, and the proposed method yields 93.4% accuracy for this database. The proposed framework provides more efficient and minimum language dependency compared to available traditional and deep learning-based approaches. The proposed model provide solution for the language independent SER model. The preprocessing of speech signal reduces the language dependency and improve the model performance. The language independent features also helps to improve the performance of multilingual SER model.

5.1 Multilingual SER Model

The block diagram of the proposed framework is presented in Fig. 5.1(a) and Fig. 5.1(b). Initially, the silence part is removed from the raw speech signal. After that, decomposition and mode tuning is performed. The noisy modes are removed at the denoising stage, and the denoised speech signal is reconstructed from the filtered modes. The relevant features are extracted from the reconstructed signal, and feature selection is performed using the rank-based algorithm. Finally, the KNN classifier is trained and tested optimal feature set to recognize different emotions.

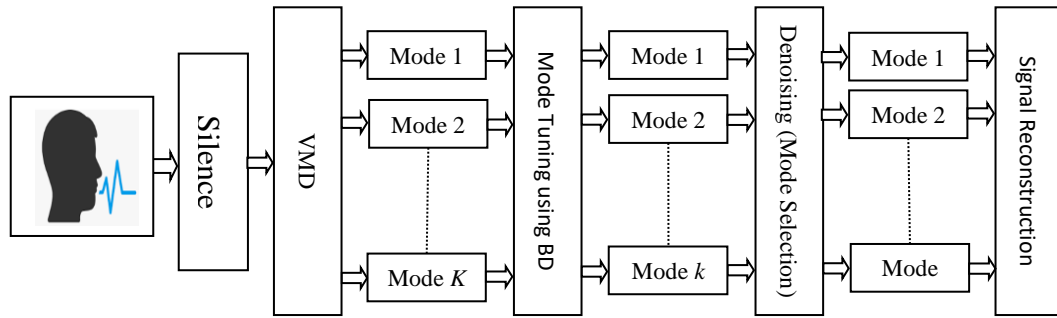


Fig. 5.1 (a) Block Diagram of Signal Reconstruction.

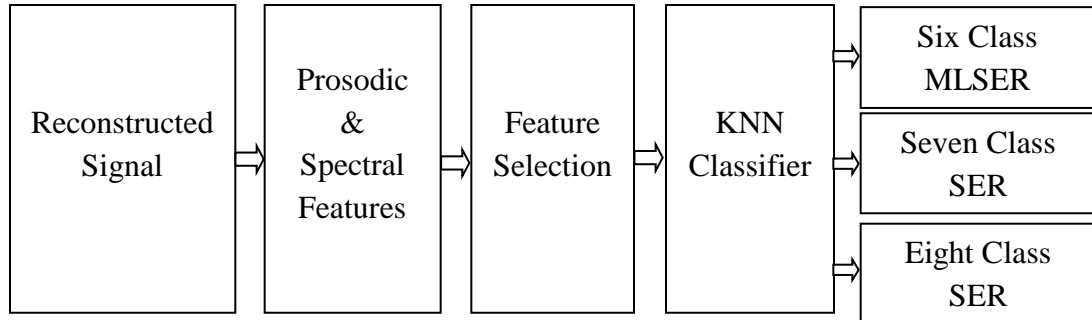


Fig. 5.1 (b) Block Diagram of Emotion Classification.

5.1.1 Silence Removal

In this framework, two simple features, short-time energy and spectral centroid, are computed to remove the silence part from the speech signal [117] the speech signal after silence removal is presented in Fig. 5.2. Let $y(n)$, $n \geq 0$ be a speech signal, and n is the discrete time. The speech signal $y(n)$ is fragmented into frames of length 50

millisecond. Let $y_l(n)$, $n=1,2,3 \dots N$ is l^{th} frame of $y(n)$. Then, for each frame short-time energy is calculated using Eqn. 5.1 [117],

$$E(l) = \frac{1}{N} \sum_{n=1}^N (y_l(n))^2 \quad (5.1)$$

The spectral centroid C_l for l^{th} frame is calculated using Eqn. 5.2 [117],

$$C_l = \frac{\sum_{r=1}^N (r+1)x_l(r)}{\sum_{r=1}^N x_l(r)} \quad r = 1, 2 \dots N. \quad (5.2)$$

Where $x_l(r)$ is the discrete Fourier transform coefficient of l^{th} frame, and N is frame length.

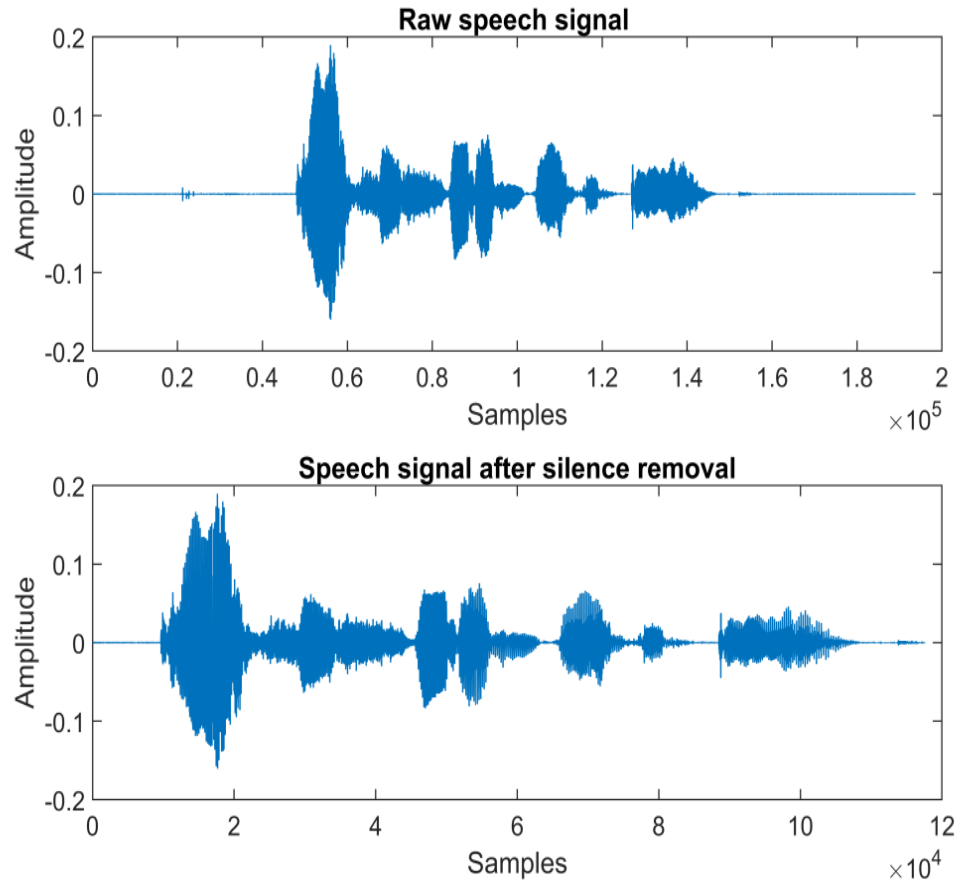


Fig. 5.2 Raw Speech Signal and Speech Signal after Silence Removal.

The extracted features are compared with threshold T_c and T_e based on the spectral centroid and energy sequences, respectively. For the calculation of T_e , the histogram of extracted feature sequence short-time energy is calculated, and a smoothing filter is applied. Then the first and second local maxima $m1$ and $m2$ are identified and used to estimate the threshold. Same process is followed for the calculation of T_c (w is a user defined parameter). The threshold value T_e is calculated using Eqn. 5.3 [117],

$$T_e = \frac{w.m1+m2}{w+1} \quad (5.3)$$

5.1.2 Noise Removal

In noise removal, mode tuning and mode selection are performed. The choice of the number of modes is based on the Improved Bhattacharyya Distance (IBD) [118]. In this method, initially, the original signal's variance and each mode's variance are computed using Eqn. 5.4 and Eqn. 5.5. After that, IBD is computed using Eqn. 5.6 and Eqn. 5.7 to measure the similarity between the original and decomposed signals. Finally, the distance between adjacent modes is calculated using Eqn. 5.8. Initially, the input signal is decomposed up to 15 modes, and IBD is calculated. It is observed that, as the number of modes increased in VMD, the similarity of the original signal with the modes is reduced, and most of the information is present in the first nine modes.

Let the A and B are two probability distributions, then variance can be defined as [118],

$$V(A) = E(A^2) + [E(A)]^2 \quad (5.4)$$

$$V(B) = E(B^2) + [E(B)]^2 \quad (5.5)$$

The Bhattacharyya distance is defined as follow [118],

$$IBD(V(A), V(B)) = -\ln[BC(V(A), V(B))] \quad (5.6)$$

Where BC is Bhattacharyya coefficient, which is defined as follow [118],

$$BC(V(A), V(B)) = \sum_{a \in A, b \in B} \sqrt{V(A)V(B)} \quad (5.7)$$

Table 5.1 Improved Bhattacharya Distance for 9 Modes and 15 Modes Decomposition of Speech Signal.

| Decomposition in 15 Modes | | | | Decomposition in 9 Modes | |
|---------------------------|-------|--------|-------|--------------------------|-------|
| Modes | IBD | Modes | IBD | Modes | IBD |
| Mode1 | 12.19 | Mode10 | 10.70 | Mode1 | 11.37 |
| Mode2 | 11.99 | Mode11 | 10.07 | Mode2 | 11.04 |
| Mode3 | 11.40 | Mode12 | 10.05 | Mode3 | 10.78 |
| Mode4 | 11.47 | Mode13 | 9.12 | Mode4 | 10.72 |
| Mode5 | 11.15 | Mode14 | 8.77 | Mode5 | 10.58 |
| Mode6 | 11.00 | Mode15 | 8.74 | Mode6 | 10.10 |
| Mode7 | 11.11 | - | - | Mode7 | 9.16 |
| Mode8 | 11.09 | - | - | Mode8 | 8.74 |
| Mode9 | 11.21 | - | - | Mode9 | 8.72 |

$$D = \max[BD(\text{Mode}(i + 1)) - BD(\text{Mode}(i))], \quad i = 1, 2, \dots, N - 1 \quad (5.8)$$

Where D is a maximum distance of adjacent mode.

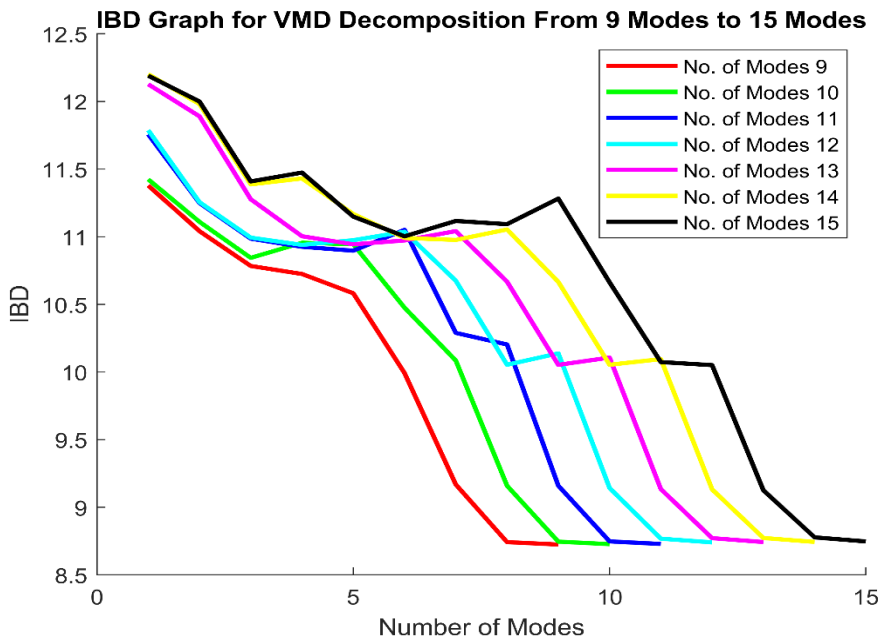


Fig. 5.3 IBD for 9 Modes to 15 Modes Decomposition.

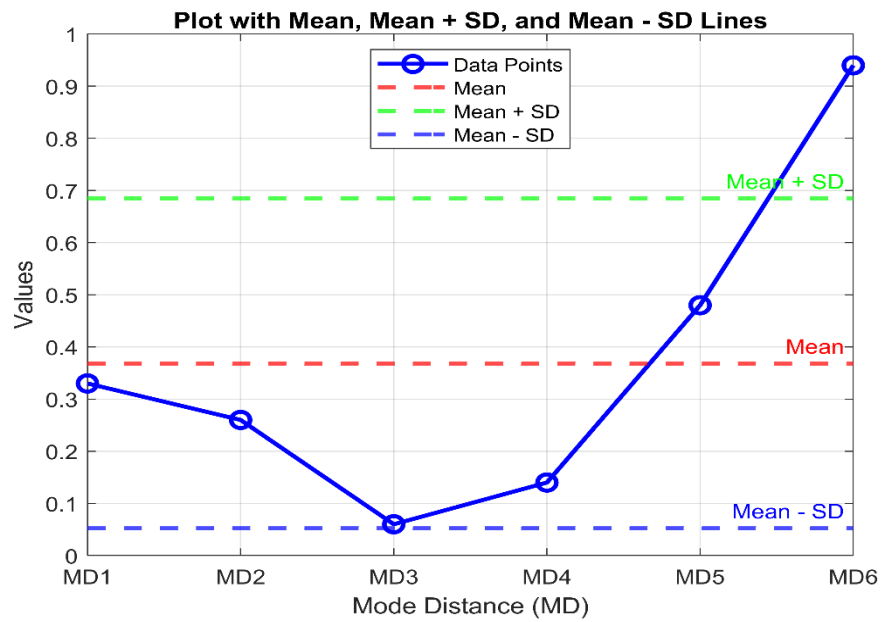


Fig. 5.4 Mean and Standard Deviation (SD) Plot for Thresholding.

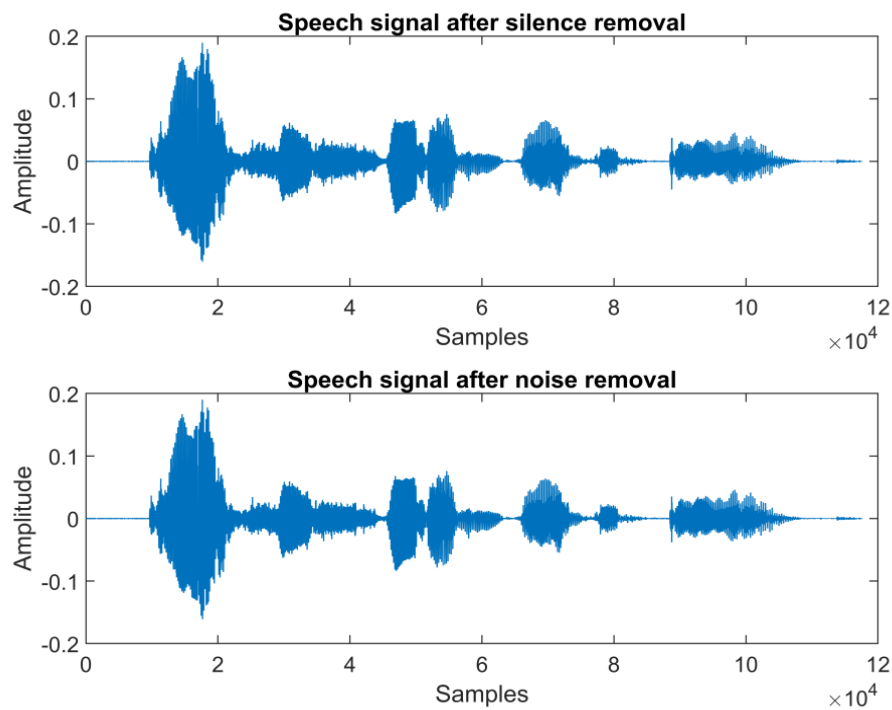


Fig. 5.5 Speech Signal after Silence Removal and Noise Removal.

Fig. 5.3 presents the graph of IBD values from 9 mode to 15 mode decomposition. Table 5.1 present an example of the IBD value of 15 mode and 9 mode decomposition. Table 5.2 signifies after the 9th mode, the IBD distance D of the adjacent mode comparatively increased. Hence, in this work signal is decomposed in nine modes only. However, it is observed in Fig. 5.4 distance between mode 6 and mode 7 (MD6) crosses the threshold value. Hence mode 7 to mode 9 are rejected and assumed last three modes contain no information. The central frequency of a VMD mode is defined as the frequency at which the mode's spectral energy is concentrated. The central frequency of each mode is calculated. The last three modes have a lower central frequency and contain noise. Finally, the signal is reconstructed using the one to six modes and residue part. After the reconstruction of the signal, it is observed that the energy of the voiced region is enhanced [119]. Fig. 5.5 shows the speech signal after silence removal and the reconstructed signal after noise removal. For the reconstruction of the signal, Eqn. 5.9 is used:

$$u = \sum_{i=3}^N mode_i + \text{residue} \quad (5.9)$$

Table 5.2 Maximum Distance of Adjacent Mode.

| Decomposition in 15 Modes | | | | Decomposition in 9 Modes | |
|---------------------------|-------|----------------|------|--------------------------|------|
| Modes | D | Modes | D | Modes | D |
| Mode1- Mode2 | 0.2 | Mode9- Mode10 | 0.51 | Mode1- Mode2 | 0.33 |
| Mode2- Mode3 | 0.59 | Mode10- Mode11 | 0.63 | Mode2- Mode3 | 0.26 |
| Mode3- Mode4 | -0.07 | Mode11- Mode12 | 0.02 | Mode3- Mode4 | 0.06 |
| Mode4- Mode5 | 0.32 | Mode12- Mode13 | 0.93 | Mode4- Mode5 | 0.14 |
| Mode5- Mode6 | 0.15 | Mode13- Mode14 | 0.35 | Mode5- Mode6 | 0.48 |
| Mode6- Mode7 | -0.11 | Mode14- Mode15 | 0.03 | Mode6- Mode7 | 0.94 |
| Mode7- Mode8 | 0.02 | - | - | Mode7- Mode8 | 0.42 |
| Mode8- Mode9 | -0.12 | - | - | Mode8- Mode9 | 0.02 |

5.1.3 Feature Extraction & Classification

The frame level features are calculated from the preprocessed signal. The reconstructed signal fragmented into 80 millisecond frame length with 75% overlapping. After that, the hamming window smoothens the edge [103]. From each frame, MFCC, mel spectrum, first and second derivative of MFCC [109], GTCC, the first derivative of GTCC, pitch, spectral crest, spectral entropy, spectral kurtosis, spectral skewness, spectral spread, harmonic ratio, and spectral centroid are calculated [4], [5].

The rank-based feature selection method ReliefF algorithm is applied to select features [110]. This algorithm selects the instances randomly from the same class and different classes. When randomly selecting R data instances from the total n instances, the ReliefF algorithm calculates the feature score f_i for each original feature ($i=1,2,3,\dots,d$), where d is the number of original features. This computation is based on Eqn. 5.10 from reference [110],

$$f_i = \frac{1}{c} \sum_{q=1}^R \left(-\frac{1}{m_q} \sum_{x_l \in NS(q)} d(X(q, i) - X(l, i)) + \sum_{z \neq q} \frac{1}{h_{qz}} \frac{p(z)}{1-p(z)} \sum_{x_l \in NH(q, z)} d(X(q, i) - X(l, i)) \right) \quad (5.10)$$

Here c represents number of classes. The term $NS(q)$ and $NH(q, y)$ are defined as the closest instances of x_q in same class of size m_q and in class z of size h_{qz} respectively. For z class the ratio of instances is defined by $p(y)$.

In the proposed framework KNN classifier is used for emotion identification. KNN classifier is based on the computation of the Euclidean distance function between pre-defined classes and each varying sample [88]. The Euclidean distance is used in the KNN algorithm to find the nearest neighbor according to each type. The most common method for calculating Euclidean distance is presented in Eqn. 5.11 [88]. Finally, the samples are assigned to the respective class based on the nearest n neighbors. In this work, fine KNN is chosen for the classification. Fine KNN takes one sample to differentiate the data. The precision, Recall, and F1-score are calculated to evaluate the performance

$$D = \sqrt{\sum_{i=1}^n (x_{1i} - x_{2i})^2} \quad (5.11)$$

Where i is the number of discrete points in each sample, and x_1, x_2 are input samples.

5.2 Results and Discussion

This work uses four database, RAVDESS speech, EMOVB, EMOVO, and multilingual database, for the experiment. In the multilingual database, the six standard classes (fear, happy, sad, neutral, angry, and disgust) of three different languages, English, German, and Italian, from three databases, RAVDESS speech, EMOVB, and EMOVO are combined. Each speech signal is preprocessed using the procedure explained in section 5.3.

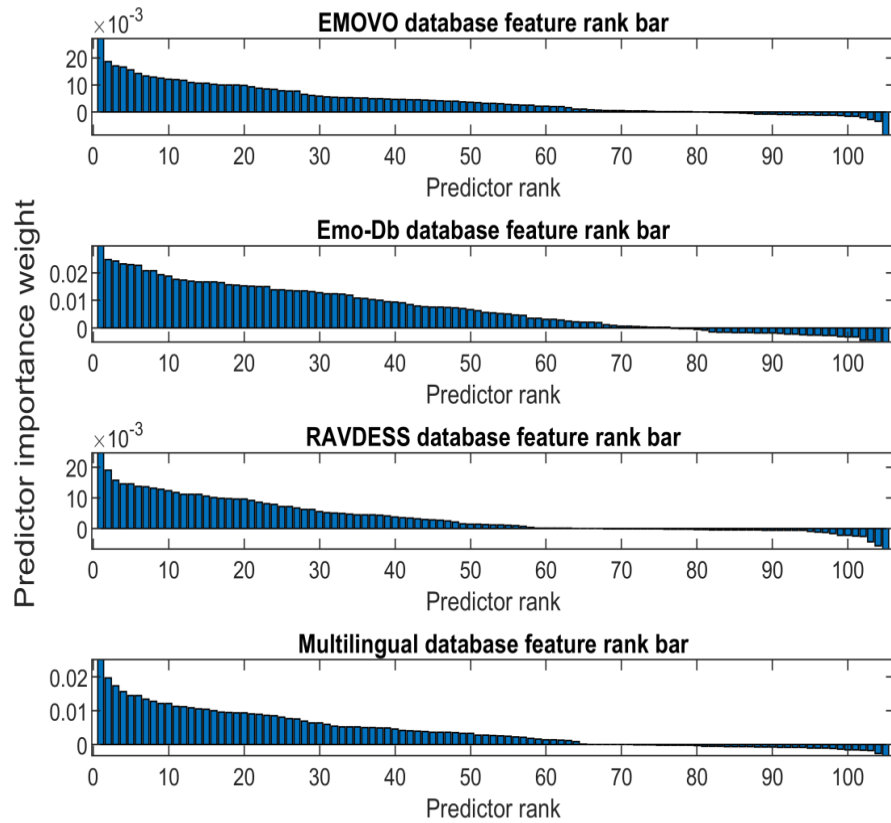


Fig. 5.6 Predictor Importance Weight Bar.

The unwanted part is removed in the preprocessing to reduce the misclassification error. After silence removal, VMD is applied to the speech signal to decompose into nine modes. The number of modes is decided using *BD*. The signal is reconstructed using the 1 to 6 modes, and the last three modes are left. The reconstructed signal is used for feature extraction. Initially, 105 prosodic and spectral features, as

explained in section 5.4, are extracted from each frame. After that, the ReliefF algorithm is used for the feature selection. The predictor importance weight bar for each database is presented in Fig. 5.6. This experiment selects features with an importance weight equal to or greater than 0.002 for the classifier training. Finally, all selected features are put into the KNN classifier for the training.

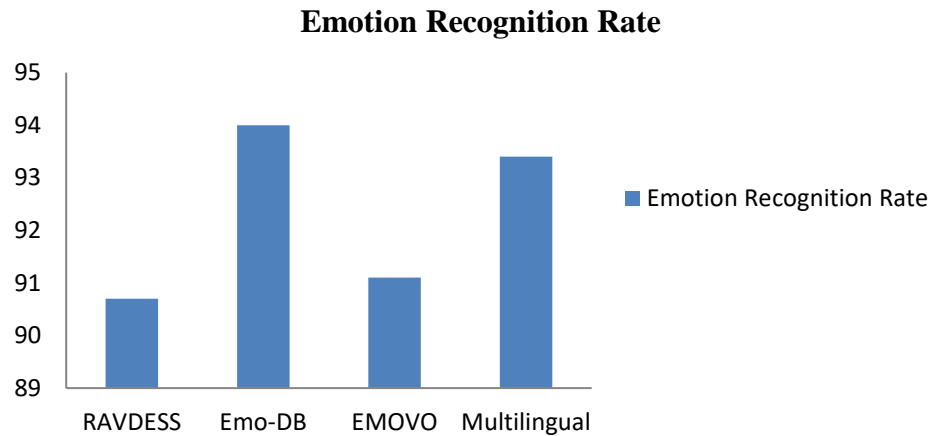


Fig. 5.7 Recognition Rate of Proposed Method for All Four Databases.

| | | | | | | | |
|---|------|------|------|------|------|------|------|
| 1 | 7423 | 256 | 57 | 100 | 82 | 67 | 51 |
| 2 | 171 | 7797 | 90 | 102 | 294 | 130 | 59 |
| 3 | 57 | 151 | 7803 | 106 | 148 | 163 | 87 |
| 4 | 127 | 115 | 83 | 6337 | 178 | 53 | 121 |
| 5 | 77 | 303 | 135 | 180 | 7091 | 63 | 96 |
| 6 | 63 | 154 | 140 | 49 | 86 | 7295 | 116 |
| 7 | 50 | 163 | 82 | 135 | 96 | 88 | 6683 |
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 |

Predicted Class

Fig. 5.8 Confusion Matrix for Italian Language Based EMOVO Database.

| | | | | | | | | |
|------------|---|-----------------|------|------|------|------|------|------|
| True Class | 1 | 10850 | 44 | 48 | 79 | 160 | 50 | 10 |
| | 2 | 16 | 8032 | 56 | 72 | 46 | 270 | 131 |
| | 3 | 37 | 51 | 4553 | 53 | 35 | 71 | 39 |
| | 4 | 42 | 75 | 52 | 5902 | 80 | 98 | 46 |
| | 5 | 141 | 63 | 44 | 104 | 6083 | 69 | 20 |
| | 6 | 32 | 285 | 64 | 99 | 45 | 7480 | 138 |
| | 7 | 1 | 123 | 35 | 46 | 7 | 92 | 4796 |
| | | Predicted Class | | | | | | |
| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 |

Fig. 5.9 Confusion Matrix for Emo-DB Database.

| | | | | | | | | | |
|------------|---|-----------------|-------|-------|-------|-------|-------|-------|-------|
| True Class | 1 | 7594 | 491 | 64 | 223 | 104 | 84 | 110 | 121 |
| | 2 | 413 | 17687 | 96 | 363 | 211 | 115 | 155 | 153 |
| | 3 | 104 | 166 | 17553 | 199 | 281 | 260 | 253 | 258 |
| | 4 | 242 | 454 | 148 | 16647 | 294 | 309 | 217 | 209 |
| | 5 | 149 | 323 | 215 | 301 | 17839 | 239 | 229 | 202 |
| | 6 | 111 | 174 | 182 | 346 | 217 | 15595 | 307 | 230 |
| | 7 | 189 | 216 | 203 | 271 | 248 | 336 | 15509 | 306 |
| | 8 | 201 | 269 | 209 | 311 | 237 | 250 | 314 | 16648 |
| | | Predicted Class | | | | | | | |
| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |

Fig. 5.10 Confusion Matrix for RAVDESS Database.

| | | | | | | |
|---|-------|-------|-------|-------|-------|-------|
| 1 | 36133 | 581 | 363 | 323 | 519 | 432 |
| 2 | 616 | 31091 | 489 | 457 | 503 | 450 |
| 3 | 355 | 330 | 28736 | 534 | 502 | 460 |
| 4 | 285 | 334 | 607 | 22954 | 287 | 370 |
| 5 | 398 | 433 | 631 | 331 | 28213 | 465 |
| 6 | 418 | 328 | 511 | 444 | 446 | 30704 |
| | 1 | 2 | 3 | 4 | 5 | 6 |

True Class

Predicted Class

Fig. 5.11 Confusion Matrix for Multilingual Database.

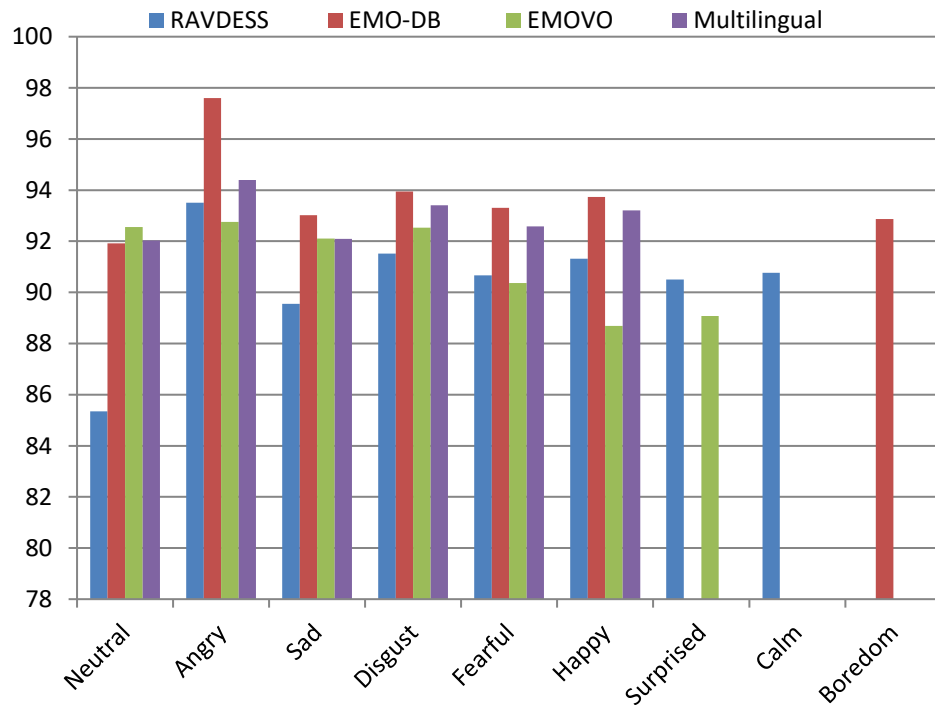


Fig. 5.12 The Class-Level Balance Accuracy for RAVDESS, Emo-DB, EMOVO and Multilingual Database.

Table 5.3 Experimental Result of Proposed Method for RAVDESS Database.

| Emotions | Precision (%) | Recall (%) | F1 score (%) |
|-----------------|----------------------|-------------------|---------------------|
| Neutral | 84.35 | 86.38 | 85.35 |
| Calm | 89.42 | 92.15 | 90.76 |
| Angry | 94.02 | 93.00 | 93.51 |
| Sad | 89.21 | 89.89 | 89.55 |
| Disgust | 91.81 | 91.24 | 91.52 |
| Fearful | 90.45 | 90.87 | 90.66 |
| Surprised | 91.26 | 89.76 | 90.50 |
| Happy | 92.35 | 90.29 | 91.31 |

Table 5.4 Experimental Result of Proposed Method for EMO-DB Database.

| Emotions | Precision (%) | Recall (%) | F1 score (%) |
|-----------------|----------------------|-------------------|---------------------|
| Angry | 98.70 | 96.52 | 97.60 |
| Boredom | 92.60 | 93.14 | 92.87 |
| Disgust | 93.83 | 94.08 | 93.95 |
| Fear | 92.87 | 93.75 | 93.31 |
| Happy | 94.22 | 93.24 | 93.73 |
| Neutral | 92 | 91.85 | 91.92 |
| Sad | 92.58 | 93.47 | 93.02 |

Table 5.5 Experimental Result of Proposed Method for EMOVO Database.

| Emotions | Precision (%) | Recall (%) | F1 score (%) |
|-----------------|----------------------|-------------------|---------------------|
| Angry | 93.16 | 92.37 | 92.76 |
| Happy | 87.22 | 90.21 | 88.69 |
| Disgust | 93.44 | 91.63 | 92.53 |
| Fear | 90.41 | 90.34 | 90.37 |
| Surprised | 88.91 | 89.25 | 89.08 |
| Neutral | 92.82 | 92.30 | 92.56 |
| Sad | 92.65 | 91.58 | 92.11 |

Table 5.6 Experimental Result of Proposed Method for Multilingual Database.

| Emotions | Precision (%) | Recall (%) | F1 score (%) |
|-----------------|----------------------|-------------------|---------------------|
| Angry | 94.57 | 94.21 | 94.39 |
| Happy | 93.93 | 92.50 | 93.21 |
| Sad | 91.69 | 92.49 | 92.09 |
| Neutral | 91.65 | 92.41 | 92.03 |
| Fearful | 92.59 | 92.58 | 92.58 |
| Disgust | 93.37 | 93.46 | 93.41 |

The accuracy of all four databases is presented in Fig. 5.7. All the classifications are performed with ten-fold cross-validation. In 10-fold cross-validation, data are split into ten subsets, where nine subsets are used for training, and one subset is used for testing. This process is repeated until each subset is used as test data. The proposed framework achieves an accuracy of 90.7% for the speech RAVDESS database, 94% for the EMO-DB database, 91.1% for the EMOVO database, and 93.4% for the multilingual database. The results presented in Fig. 5.7 show that the proposed framework has the highest sensitivity for the EMO-DB database and the lowest sensitivity for the speech RAVDESS database. The confusion matrix of all four databases is presented from

Fig. 5.8 to Fig. 5.11 (Emotions serial number is correspond to respective Precision, Recall and F1 Score table), and class-level balance accuracy is illustrated in Fig. 5.12.

The precision, recall, and F1-score for each emotion are presented in Table 5.3 to Table 5.6. For the RAVDESS speech database, the highest precision of 94.02% is obtained for angry mode, while the highest recall of 93.0% and 92.15% are achieved for angry and calm mode, respectively. For the EMO-DB speech database highest precision, 98.7%, is obtained for angry mode, and the highest recall, 96.52%, is achieved for angry mode. Similarly, For the EMOVO speech database, the highest precision of 93.44% is obtained for disgust mode, and the highest recall of 92.37% is achieved for angry mode. Finally, in Table 5.6, the proposed architecture is tested for the multilingual database. For the multilingual speech database highest precision, 94.57%, is obtained for angry mode, and the highest recall, 94.21%, is achieved for angry mode. Table 5.3 to Table 5.6 show that the proposed emotion recognition architecture is comparatively more sensitive to angry class emotion.

In Table 5.7 and Table 5.8, the proposed framework is compared with the other existing methods. Table 5.7 shows the emotion recognition accuracy of the individual database, and Table 5.8 shows the emotion recognition accuracy of multilingual database. Table 5.7 shows that all the available existing methods are highly language dependent. In [93], the Italian language database accuracy is 91.16%, while for the German language-based database, the achieved accuracy is 86.22% and shows approximately 5% variation in accuracy with the language change. In [113], the English language database accuracy is 85.57%, while for the German language-based database, the achieved accuracy is 77.02%. This method shows more than 8% variation in accuracy with the language change.

Similarly, [96] shows more than 24%, [98] shows more than 17%, [99] shows 10%, and [71] shows more than 10% variation in accuracy and shows high language dependency. Compared to existing methods, the proposed method has only 3.3% variation and proved more robust and has minimum language dependency. Many ways are available for the individual database, but more research is needed for the multilingual database. Because of the diverse language, the accuracy is drastically reduced. For example, in [120], the highest achieved accuracy for the German language was 90.09%, but it declined to 80.05% for the multilingual database. This signifies the existing frameworks are highly language dependent. In contrast, the proposed method shows a low language dependency hence the highest achieved accuracy for the German language is 94%, and for the multilingual is 93.4%. Only German, English, and Italian language-based research work is included for the multilingual database comparison. The proposed method achieves the highest accuracy for the diverse language database compared to all existing methods.

Table 5.7 Performance Comparison Benchmark for Individual Database.

| Author | Database | Method | Accuracy (%) | Language Sensitivity (%) |
|----------------------------|--|---|----------------------------------|--------------------------|
| Kerkeni, et al. [93] | EMOVO EMO-DB | EMD combined with the Teager-Kaiser Energy Operator and then features are extracted. | 91.16 86.22 | 4.94 |
| M. Mustaqeem, et al. [113] | EMO-DB RAVDESS Speech | Sequence based spectrogram is passed through CNN to extract feature. | 85.57 77.02 | 8.55 |
| M. Mustaqeem, et al. [77] | RAVDESS Speech | Spectrogram of speech signals is passed through the deep stride convolutional neural network (DSCNN). | 79.5 | - |
| T. Özseven, et al. [96] | EMO-DB SAVEE EMOVO | Emotion based acoustic feature selection method is used. | 84.62 74.39 60.4 | 24.22 |
| J.Ancilin, et al. [98] | EMO-DB RAVDESS Speech EMOVO | Energy spectrum and Mel Frequency Magnitude Coefficient based features are used. | 81.5 64.31 73.3 | 17.19 |
| Kwon, et al. [99] | EMO-DB SAVEE RAVDESS Speech | Iterative Neighborhood Component Analysis (INCA) is applied for learning spatial-spectral features that are mutually learned, and then the optimal features are selected. | 95 82 85 | 13 |
| Tuncer, et al. [71] | EMOVO RAVDESS Speech EMO-DB SAVEE | Used twine shuffle pattern with Tunable Q wavelet transform for multi-level feature generation. | 90.09 84.79 79.08 87.43 | 11.01 |
| Proposed Method | RAVDESS Speech EMO-DB EMOVO | VMD decomposition and IBD based mode tuning and denoising after that features are extracted. | 90.70 94 91.10 | 3.30 |

Table 5.8 Performance Comparison Benchmark for Multilingual Database.

| Author | language | Classifier | Method | Accuracy (%) |
|------------------------|---------------------------------------|-----------------|--|--------------|
| Tuncer, et al. [71] | German English Italian | SVM | Used twine shuffle pattern with Tunable Q wavelet transform for multi-level feature generation. | 80.05 |
| Proposed Method | German English Italian | Fine KNN | Used VMD decomposition, BD based mode tuning and denoising after that features are extracted. | 93.4 |

5.3 Summary

The proposed framework introduced a decomposition-based denoising approach for MLSER. The method includes the steps of silence removal, signal decomposition, number of mode selections, denoising and signal reconstruction. These steps make the signal more predictive for the recognition of respective emotions. After that, spectral and prosodic features are extracted, and the ReliefF algorithm is used to explore the optimal feature set. This optimal feature set is put into the KNN classifier for training and testing. The proposed method obtained 90.70%, 94%, and 91.10% accuracies for RAVDESS Speech, EMO-DB, and EMOVO database, respectively. For multilingual database, the proposed method obtained 93.4% accuracy. The performance comparison shows that the proposed method provides more stable performance on multiple databases and the lowest language dependency, basically the needs of an ideal MLSER architecture.

CHAPTER 6

CONCLUSION, FUTURE SCOPE AND SOCIAL IMPACT

6.1 Conclusion

This thesis work provides a signal processing framework to enhance speech signal predictability for emotion recognition. It is observed that data-driven methods like VMD are more suitable for speech signal preprocessing compared to other decomposition methods. Additionally, this work offers an effective SER model for binary class, multiclass, and multilingual emotion identification. The performance of the proposed techniques is validated on benchmark datasets, demonstrating significant improvements over traditional approaches.

In this thesis work, first RDWT based method is proposed. In this work RDWT decomposes speech signals into 8 SBs per frame. Features such as complexity, AAC, mobility, and ZCR are extracted from each SB and their statistical significance is evaluated using the p-value of the KW test. These features are used for training and testing multiple classifier variants. The KNN and ensemble variants demonstrated superior performance compared to DT variants. Among all classifiers, the OE classifier achieved the highest accuracy of 83.3%, outperforming existing works in this domain.

Expanding the methodology, an EMD-based approach is introduced for feature extraction and classification, achieving superior accuracy. The EMD based method decomposes input speech signals into IMFs and extracts features such as the energy-based ratio and MFCC from each IMF. The statistical significance of these features is assessed using the ReliefF algorithm and selected higher rank features for further analysis. The EMD-based method leverages OE classifiers with 10-fold cross-validation, achieving an accuracy of 95.3% for binary classification, outperforming state-of-the-art techniques.

To further improve performance, the proposed SER architecture leverages VMD and TKEO for robust signal analysis and classification across various emotion categories. The proposed SER architecture was tested using the RAVDESS dataset. Features extracted from the VMD-TKEO processed signals are statistically examined using the p -value of the KW test. The selected feature set is evaluated using variants of the SVM classifier for two-class, three-class, and four-class emotion categories. The QSVM provided the best results across all categories, with recognition rates of 100%, 92.2%, and 81.7% for two-class, three-class, and four-class categories, respectively.

Despite SER performance improvement still there remains scope for improving multiclass SER system's performance. The proposed method demonstrates that VMD decomposition and dominant mode selection enhance signal predictability. Experimental results indicate that most informative features reside in dominant modes, while lower energy modes contain less useful information. The feature optimization algorithm helps to select the potential feature set for emotion classification. To reduce computational complexity, the KNN classifier is employed, as it does not involve weight computation or parameter adjustment, making it computationally efficient. The method's robustness is verified on four datasets: RAVDESS-speech, Emo-DB, EMOVO, and IEMOCAP proving effective across three languages: English, German, and Italian.

Furthermore, a decomposition-based approach is proposed for MLSER. This approach includes steps such as silence removal, signal decomposition, mode selection, denoising, and signal reconstruction, which enhance signal predictability for emotion recognition. After these steps, spectral and prosodic features are extracted, and the ReliefF algorithm is used to identify the optimal feature set, which is then used to train and test the KNN classifier. The proposed method achieved accuracies of 90.70%, 94%, and 91.10% for the RAVDESS Speech, EMO-DB, and EMOVO databases, respectively. For the multilingual database, the proposed method achieved 93.4% accuracy. Performance comparisons indicate that the proposed method offers more stable performance across multiple databases with minimal language dependency, aligning with the requirements of an ideal MLSER architecture. The performance was benchmarked against other state-of-the-art methods, showcasing its superiority in handling diverse datasets. These results highlight the robustness and scalability of the method, emphasizing its potential for deployment in practical multilingual emotion recognition systems.

6.2 Future Scope

While thesis has made significant contributions to SER by introducing novel methods but still there remains substantial potential for future exploration and

enhancement. There is also potential for expanding the current framework to support more languages, especially those with different phonetic and tonal characteristics. While the proposed method has demonstrated efficacy across three languages, developing a truly language independent SER system would require advanced cross-lingual feature transfer and domain adaptation techniques. Furthermore, real-time implementation of these methods could be explored, as current processes, particularly in signal decomposition and feature extraction, are computationally intensive. Additionally, combining various feature extraction techniques into hybrid feature sets may yield a more comprehensive representation of speech signals, enhancing the robustness of emotion recognition systems across diverse datasets. Improving the robustness of SER systems in noisy, real-world environments is another critical area for future work. While the denoising approaches introduced in this research are effective, further refinements could ensure stable performance in practical applications where background noise is prevalent. Finally, as SER systems become more complex, the need for interpretability and explainability will grow, especially in sensitive applications such as mental health diagnostics. Future research could explore methods to make machine learning models in SER more transparent, providing clear explanations for their emotional classifications and decisions.

6.3 Social Impact

The advancements in emotion recognition through speech signals presented in this thesis hold profound social implications, addressing critical needs across language diversity, accessibility, and social integration. This research provides lower language sensitivity; hence, a common model can be used for different languages. It enhances human-computer interaction, making technology more intuitive and responsive to human emotions, which can improve user experience and accessibility. For individuals with cognitive impairments, this technology offers significant benefits by providing tools that can better understand and respond to their emotional states, aiding in communication and social integration. Moreover, applications in mental health monitoring and support can offer timely interventions and personalized care, contributing to overall well-being and quality of life.

REFERENCES

- [1] Cowie, Roddy, Ellen Douglas-Cowie, Nicolas Tsapatsoulis, George Votsis, Stefanos Kollias, Winfried Fellenz, and John G. Taylor, "Emotion recognition in human-computer interaction," *IEEE Signal processing magazine*, vol. 18, pp. 32-80, 2001.
- [2] Zeng, Zhihong, Maja Pantic, Glenn I. Roisman, and Thomas S. Huang, "A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, pp. 39-58, 2009.
- [3] Scherer, Klaus R, "What are emotions? And how can they be measured?," *Social science information*, vol. 44, pp. 695-729, 2005.
- [4] Ververidis, Dimitrios and Kotropoulos, Constantine, "Emotional speech recognition: Resources, features, and methods," *Speech communication*, vol. 48, pp. 1162-1181, 2006.
- [5] El Ayadi, Moataz and Kamel, Mohamed S and Karray, Fakhri, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern recognition*, vol. 44, pp. 572-587, 2011.
- [6] Georgogiannis, Alexandros and Digalakis, Vassilis, "Speech emotion recognition using non-linear teager energy based features in noisy environments," in *proceedings of the 20th European signal processing conference (EUSIPCO)*, 2012.
- [7] Kragel, Philip A and LaBar, Kevin S, "Multivariate pattern classification reveals autonomic and experiential representations of discrete emotions," *Emotion*, vol. 13, pp. 681-690, 2013.
- [8] Tawari, Ashish and Trivedi, Mohan Manubhai, "Speech emotion analysis: Exploring the role of context," *IEEE Transactions on multimedia*, vol. 12, pp. 502-509, 2010.
- [9] Lokesh, S and Devi, M Ramya, "Speech recognition system using enhanced mel frequency cepstral coefficient with windowing and framing method," *Cluster Computing*, vol. 22, pp. 11669-11679, 2019.
- [10] Christiansen, Claus and Pedersen, Michael Syskind and Dau, Torsten, "Prediction of speech intelligibility based on an auditory preprocessing model," *Speech Communication*, vol. 52, pp. 678-692, 2010.
- [11] Liu, Man, "English speech emotion recognition method based on speech recognition," *International Journal of Speech Technology*, vol. 25, pp. 391-398, 2022.
- [12] Al-Hashemy, BAR and Taha, SMR, "Voiced-unvoiced-silence classification of speech signals based on statistical approaches," *Applied Acoustics*, vol. 25, pp. 169-179, 1988.
- [13] Tsanas, Athanasios and Little, Max A and McSharry, Patrick E and Spielman, Jennifer and Ramig, Lorraine O, "Novel speech signal processing algorithms for high-accuracy classification of Parkinson's disease," *IEEE transactions on biomedical engineering*, vol. 59, pp. 1264-1271, 2012.
- [14] Pan, Bei and Hirota, Kaoru and Jia, Zhiyang and Dai, Yaping, "A review of multimodal emotion recognition from datasets, preprocessing, features, and fusion methods," *Neurocomputing*, vol. 561, p. 126866, 2023.

- [15] Chun-Chieh Wang, Yuan Kang, "Feature Extraction Techniques of Non-Stationary Signals for Fault Diagnosis in Machinery Systems," *Journal of Signal and Information Processing*, vol. 3, pp. 2159-4465, 2012.
- [16] Nayak, Maya and Panigrahi, Bhawani Sankar, "Advanced signal processing techniques for feature extraction in data mining," *International Journal of Computer Applications*, vol. 19, pp. 30-37, 2011.
- [17] R. Fonseca-Pinto, "A New Tool for Nonstationary and Nonlinear Signals: The Hilbert-Huang Transform in Biomedical Applications," in *Biomedical Engineering, Trends in Electronics, Communications and Software*, 2011, pp. 481-502.
- [18] Mohammadi, Zeynab and Frounchi, Javad and Amiri, Mahmood, "Wavelet-based emotion recognition system using EEG signal," *Neural Computing and Applications*, vol. 28, pp. 1985-1990, 2017.
- [19] Li, Xiang and Li, Xin and Zheng, Xiaoming and Zhang, Dexing, "EMD-TEO based speech emotion recognition," in *International Conference on Intelligent Computing for Sustainable Energy and Environment*, 2010.
- [20] Shahnaz, Celia and Sultana, Sharifa and Fattah, Shaikh Anowarul and Rafi, RH Md and Ahmmmed, Istak and Zhu, W-P and Ahmad, M Omair, "Emotion recognition based on EMD-Wavelet analysis of speech signals," in *IEEE International Conference on Digital Signal Processing (DSP)*, 2015.
- [21] Zhuang, Ning and Zeng, Ying and Tong, Li and Zhang, Chi and Zhang, Hanming and Yan, Bin, "Emotion recognition from EEG signals using multidimensional information in EMD domain," *BioMed research international*, vol. 17, pp. 8317357, 2017.
- [22] Riaz, Farhan and Hassan, Ali and Rehman, Saad and Niazi, Imran Khan and Dremstrup, Kim, "EMD-based temporal and spectral features for the classification of EEG signals using supervised learning," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 24, pp. 28-35, 2015.
- [23] Gupta, Akshansh and Kumar, Dharendra and Verma, Hanuman and Tanveer, Muhammad and Javier, Andreu Perez and Lin, Chin-Teng and Prasad, Mukesh, "Recognition of multi-cognitive tasks from EEG signals using EMD methods," *Neural Computing and Application*, vol. 35, pp. 22989-23006, 2023.
- [24] Norden, E Huang and Zheng, Shen and Steven, R Long and Manli, C Wu and Hsing, H Shih and Quanan, Zheng and Nai-Chyuan, Yen and Chi, Chao Tung and Henry, H Liu, "The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis," *Proc. R. Soc. Lond. A*, vol. 454, pp. 903-995, 1998.
- [25] Dragomiretskiy, Konstantin and Zosso, Dominique, "Variational mode decomposition," *IEEE transactions on signal processing*, vol. 62, pp. 531-544, 2013.
- [26] Nagineni, Sukumar and Taran, Sachin and Bajaj, Varu, "Features based on variational mode decomposition for identification of neuromuscular disorder using EMG signals," *Health information science and systems*, vol. 6, pp. 1-10, 2018.
- [27] Lahmiri, Salim and Boukadoum, Mounir, "Physiological signal denoising with variational mode decomposition and weighted reconstruction after DWT thresholding," in *IEEE international symposium on circuits and systems (ISCAS)*, 2015.
- [28] Hermes, Dik J, "Measurement of pitch by subharmonic summation," *The journal of the acoustical society of America*, vol. 83, pp. 257-264, 1988.

- [29] Schuller, B., Rigoll, G. and Lang, M., "Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture," in *IEEE international conference on acoustics, speech, and signal processing*, 2004.
- [30] Lugger, Marko and Yang, Bin, "The relevance of voice quality features in speaker independent emotion recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, 2007.
- [31] Nwe, Tin Lay and Foo, Say Wei and De Silva, Liyanage C, "Speech emotion recognition using hidden Markov models," *Speech communication*, vol. 41, pp. 603-623, 2003.
- [32] Palo, HK and Mohanty, Mihir Narayana and Chandra, Mahesh, "Use of different features for emotion recognition using MLP network," in *Computational Vision and Robotics: Proceedings of ICCVR*, 2015.
- [33] Yogesh, CK and Hariharan, Muthusamy and Yuvaraj, R and Ngadiran, Ruzelita and Yaacob, Sazali and Polat, Kemal and others, "Bispectral features and mean shift clustering for stress and emotion recognition from natural speech," *Computers & Electrical Engineering*, vol. 62, pp. 676-691, 2017.
- [34] Teager, H, "Some observations on oral air flow during phonation," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, pp. 599-601, 1980.
- [35] Cairns, Douglas A and Hansen, John HL, "Nonlinear analysis and classification of speech under stressed conditions," *The Journal of the Acoustical Society of America*, vol. 96, pp. 3392-3400, 1994.
- [36] Wu, Kebin and Zhang, David and Lu, Guangming, "GMAT: Glottal closure instants detection based on the multiresolution absolute Teager--Kaiser energy operator," *Digital Signal Processing*, vol. 69, pp. 286-299, 2017.
- [37] Sun, Rui and Moore, Elliot, "Investigating glottal parameters and teager energy operators in emotion recognition," in *International Conference on Affective Computing and Intelligent Interaction*, 2011.
- [38] Bou-Ghazale, Sahar E and Hansen, John HL, "A comparative study of traditional and newly proposed features for recognition of speech under stress," *IEEE Transactions on speech and audio processing*, vol. 8, pp. 429-442, 2000.
- [39] Alpan, Ali and Schoentgen, Jean and Maryn, Youri and Grenez, Francis, "Automatic perceptual categorization of disordered connected speech," in *11th Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2010.
- [40] Attabi, Yazid, Md Jahangir Alam, Pierre Dumouchel, Patrick Kenny, and Douglas O'Shaughnessy, "Multiple windowed spectral features for emotion recognition," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013.
- [41] Wang, Yan and Hu, Weiping, "Speech emotion recognition based on improved MFCC," in *Proceedings of the 2nd international conference on computer science and application engineering*, 2018.
- [42] Bou-Ghazale, Sahar E and Hansen, John HL, "A comparative study of traditional and newly proposed features for recognition of speech under stress," *IEEE Transactions on speech and audio processing*, vol. 8, pp. 429-442, 2000.

- [43] Liu, Zhen-Tao and Xie, Qiao and Wu, Min and Cao, Wei-Hua and Li, Dan-Yun and Li, Si-Han, "Electroencephalogram emotion recognition based on empirical mode decomposition and optimal feature selection," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 11, pp. 517-526, 2018.
- [44] Yang, Bin and Lugger, Marko, "Emotion recognition from speech signals using new harmony features," *Signal processing*, vol. 90, pp. 1415-1423, 2010.
- [45] Sun, Yaxin and Wen, Guihua and Wang, Jiabing, "Weighted spectral features based on local Hu moments for speech emotion recognition," *Biomedical signal processing and control*, vol. 18, pp. 80-90, 2015.
- [46] Ying, Sun and Xue-Ying, Zhang, "Characteristics of human auditory model based on compensation of glottal features in speech emotion recognition," *Future Generation Computer Systems*, vol. 81, pp. 291-296, 2018.
- [47] Teager, HM and Teager, SM, "Evidence for nonlinear sound production mechanisms in the vocal tract," *Speech production and speech modelling*, pp. 241-261, 1990.
- [48] Kaiser, James F, "On a simple algorithm to calculate the 'energy' of a signal," in *International conference on acoustics, speech, and signal processing*, 1990.
- [49] Zhou, Guojun and Hansen, John HL and Kaiser, James F}, "Nonlinear feature based classification of speech under stress," *IEEE Transactions on speech and audio processing*, vol. 9, pp. 201-216, 2001.
- [50] Liu, Zhen-Tao and Xie, Qiao and Wu, Min and Cao, Wei-Hua and Mei, Ying and Mao, Jun-Wei, "Speech emotion recognition based on an improved brain emotion learning model," *Neurocomputing*, vol. 309, pp. 145-156, 2018.
- [51] Schuller, Björn, Stefan Steidl, and Anton Batliner, "The INTERSPEECH 2009 emotion challenge," in *Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2009.
- [52] Schuller, Björn, Stefan Steidl, Anton Batliner, Felix Burkhardt, Laurence Devillers, Christian Müller, and Shrikanth Narayanan, "The INTERSPEECH 2010 paralinguistic challenge,," in *11th Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2010.
- [53] Khalid, Samina and Khalil, Tehmina and Nasreen, Shamila, "A survey of feature selection and feature extraction techniques in machine learning," in *science and information conference*, 2014.
- [54] "Singular value decomposition and principal component analysis," in *Wall, Michael E and Rechtsteiner, Andreas and Rocha, Luis M*, Springer, 2003, pp. 91-109.
- [55] Bartenhagen, Christoph and Klein, Hans-Ulrich and Ruckert, Christian and Jiang, Xiaoyi and Dugas, Martin, "Comparative study of unsupervised dimension reduction techniques for the visualization of microarray gene expression data," *BMC bioinformatics*, vol. 11, pp. 1-11, 2010.
- [56] Ang, Jun Chin and Mirzal, Andri and Haron, Habibollah and Hamed, Haza Nuzly Abdull, "Supervised, unsupervised, and semi-supervised feature selection: a review on gene selection," *IEEE/ACM transactions on computational biology and bioinformatics*, vol. 13, pp. 971-989, 2015.
- [57] Ang, Jun Chin and Mirzal, Andri and Haron, Habibollah and Hamed, Haza Nuzly Abdull, "Supervised, unsupervised, and semi-supervised feature selection: a review on gene

- selection," *IEEE/ACM transactions on computational biology and bioinformatics*, vol. 13, pp. 971-989, 2015.
- [58] Chen, Hao and Chen, Hongmei and Li, Weiyi and Li, Tianrui, "Semi-supervised feature selection based on pairwise constraint-guided dual space latent representation learning and double sparse graphs discriminant," *Applied Intelligence*, vol. 53, pp. 12288-12307, 2023.
 - [59] Cherrington, Marianne and Thabtah, Fadi and Lu, Joan and Xu, Qiang, "Feature selection: filter methods performance challenges," in *International Conference on Computer and Information Sciences (ICCIS)*, 2019.
 - [60] Chen, Gang and Chen, Jin, "A novel wrapper method for feature selection and its applications," *Neurocomputing*, vol. 159, pp. 219-226, 2015.
 - [61] Liu, Haoyue and Zhou, MengChu and Liu, Qing, "An embedded feature selection method for imbalanced data classification," *IEEE/CAA Journal of Automatica Sinica*, vol. 6, pp. 703-715, 2019.
 - [62] Wei, Guangfen and Zhao, Jie and Feng, Yanli and He, Aixiang and Yu, Jun, "A novel hybrid feature selection method based on dynamic feature importance," *Applied Soft Computing*, vol. 93, p. 106337, 2020.
 - [63] Ben Brahim, Afef and Limam, Mohamed, "Ensemble feature selection for high dimensional data: a new method and a comparative study," *Advances in Data Analysis and Classification*, vol. 12, pp. 937-952, 2018.
 - [64] Jingjie, YAN and Xiaolan, WANG and Weiyi, GU and LiLi, MA, "Speech emotion recognition based on sparse representation," *Archives of Acoustics*, vol. 2013, pp. 465-470, 38.
 - [65] Yu, Wenbo and Zhang, Miao and Shen, Yi, "Spatial revising variational autoencoder-based feature extraction method for hyperspectral images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, pp. 1410-1423, 2020.
 - [66] Siddique Latif, Rajib Rana, Junaid Qadir, and Julien Epps, "Variational autoencoders to learn latent representations of speech emotion," in *Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2018.
 - [67] Lee, Chul Min and Narayanan, Shrikanth, "Toward detecting emotions in spoken dialogs," *IEEE transactions on speech and audio processing*, vol. 13, pp. 293-303, 2005.
 - [68] Wu, Dongrui, Thomas D. Parsons, and Shrikanth Narayanan, "Acoustic feature analysis in speech emotion primitives estimation," in *Proceedings of the 11th Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2010.
 - [69] Landgrebe, Thomas CW and Duin, Robert PW, "Efficient multiclass ROC approximation by decomposition via confusion matrix perturbation analysis," *IEEE transactions on pattern analysis and machine intelligence*, vol. 30, pp. 810-822, 2008.
 - [70] Das, Resul, "A comparison of multiple classification methods for diagnosis of Parkinson disease," *Expert Systems with Applications*, vol. 37, pp. 1568-1572, 2010.
 - [71] Tuncer, Turker, Sengul Dogan, and U. Rajendra Acharya, "Automated accurate speech emotion recognition system using twine shuffle pattern and iterative neighborhood component analysis techniques," *Knowledge Based Systems*, vol. 211, p. 106547, 2021.
 - [72] Jannat, Rahatul and Tynes, Iyonna and Lime, Lott La and Adorno, Juan and Canavan, Shaun, "Ubiquitous emotion recognition using audio and video data," in *Proceedings of*

the 2018 ACM international joint conference and 2018 International symposium on pervasive and ubiquitous computing and wearable computers, 2018.

- [73] Spyrou, Evaggelos, Rozalia Nikopoulou, Ioannis Vernikos, and Phivos Mylonas, "Emotion recognition from speech using the bag-of-visual words on audio segment spectrograms," *Technologies*, vol. 7, pp. 20, 2019.
- [74] Jalal, Md Asif and Loweimi, Erfan and Moore, Roger K and Hain, Thomas, "Learning temporal clusters using capsule routing for speech emotion recognition," in *Proceedings of interspeech*, 2019.
- [75] Krishnan, Palani Thanaraj, Alex Noel Joseph Raj, and Vijayarajan Rajangam, "Emotion classification from speech signal based on empirical mode decomposition and non-linear features: Speech emotion recognition," *Complex & Intelligent Systems*, vol. 7, pp. 1919-1934, 2021.
- [76] Livingstone, Steven R and Russo, Frank A, "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English," *PloS one, Public Library of Science San Francisco, CA USA*, vol. 13, pp. e0196391, 2018.
- [77] Mustaqeem and Kwon, Soonil, "A CNN-assisted enhanced audio signal processing for speech emotion recognition," *Sensors*, vol. 20, pp. 183-191, 2019.
- [78] Dutt, Aditya, and Paul Gader, "Wavelet multiresolution analysis based speech emotion recognition system using 1D CNN LSTM networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 2043-2054, 2023.
- [79] Prakash, Chandra and Gaikwad, VB and Singh, Ravish R and Prakash, Om, "Analysis of emotion recognition system through speech signal using KNN \& GMM classifier," *IOSR J. Electron. Commun. Eng.(IOSR-JECE)*, vol. 10, pp. 55-61, 2015.
- [80] Emerich, Simina and Lupu, Eugen and Apatean, Anca, "Emotions recognition by speechand facial expressions analysis," in *17th European signal processing conference*, 2009.
- [81] Deb, Suman, Samarendra Dandapat, and Jarek Krajewski, "Analysis and Classification of Cold Speech Using Variational Mode Decomposition," *IEEE Transactions on Affective Computing, IEEE*, vol. 11, pp. 296-307, 2020.
- [82] Ziqiang Liu, Jingjia Jia, and Wensheng Sun, "CASIA corpus: A Chinese emotional speech corpus," in *Int. Conf. Chinese Spoken Lang. Process. (ISCSLP)*, Singapore, 2006.
- [83] Chen, Bu and Yin, Qian and Guo, Ping, "A study of deep belief network based Chinese speech emotion recognition," in *Tenth International Conference on Computational Intelligence and Security*, 2014.
- [84] Huang, Jian and Liu, Bin and Tao, Jianhua, "Learning long-term temporal contexts using skip RNN for continuous emotion recognition," *Virtual Reality and Intelligent Hardware*, vol. 3, pp. 55-64, 2021.
- [85] Zhang, Weishan and Zhao, Dehai and Chai, Zhi and Yang, Laurence T and Liu, Xin and Gong, Faming and Yang, Su, "Deep learning and SVM-based emotion recognition from Chinese speech for smart affective services," *Software: Practice and Experience*, vol. 47, pp. 1127-1138, 2017.

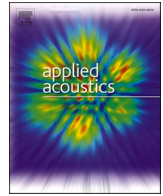
- [86] Prasomphan, Sathit, "Improvement of speech emotion recognition with neural network classifier by using speech spectrogram," in *International Conference on Systems, Signals and Image Processing (IWSSIP)*, 2015.
- [87] Han, Zhongyang and Zhao, Jun and Leung, Henry and Ma, King Fai and Wang, Wei, "A review of deep learning models for time series prediction," *IEEE Sensors Journal*, vol. 21, pp. 7833-7848, 2019.
- [88] Bombatkar, Anuja and Bhoyar, Gayatri and Morjani, Khushbu and Gautam, Shalaka and Gupta, Vikas, "Emotion recognition using Speech Processing Using k-nearest neighbor algorithm," *Int. J. Eng. Res. Appl*, vol. 4, pp. 68-71, 2014.
- [89] Zhang, Biqiao and Provost, Emily Mower and Essl, Georg, "Cross-corpus acoustic emotion recognition from singing and speaking: A multi-task learning approach," in *IEEE International conference on acoustics, speech and signal processing (ICASSP)*, 2016.
- [90] Danisman, Taner and Alpkocak, Adil, "Emotion classification of audio signals using ensemble of support vector machines," in *International tutorial and research workshop on perception and interactive technologies for speech-based systems*, Springer, 2008, pp. 205-216.
- [91] Liu, Zhen-Tao, Min Wu, Wei-Hua Cao, Jun-Wei Mao, Jian-Ping Xu, and Guan-Zheng Tan, "Speech emotion recognition based on feature selection and extreme learning machine decision tree," *Neurocomputing*, vol. 273, pp. 271-280, 2018.
- [92] Lalitha, S., Deepa Gupta, Mohammed Zakariah, and Yousef Ajami Alotaibi, "Investigation of multilingual and mixed-lingual emotion recognition using enhanced cues with data augmentation," *Applied Acoustics*, vol. 170, pp. 107519, 220.
- [93] Kerkeni, Leila, Youssef Serrestou, Kosai Raoof, Mohamed Mbarki, Mohamed Ali Mahjoub, and Catherine Cleder, "Automatic speech emotion recognition using an optimal combination of features based on EMD-TKEO," *Speech communication, Elsevier*, vol. 114, pp. 22-35, 2019.
- [94] Ezz-Eldin, Mai, Ashraf AM Khalaf, Hesham FA Hamed, and Aziza I. Hussein, "Efficient feature-aware hybrid model of deep learning architectures for speech emotion recognition," *IEEE Access, IEEE*, vol. 9, pp. 19999-20011, 2021.
- [95] Xie, Yue, Ruiyu Liang, Zhenlin Liang, Chengwei Huang, Cairong Zou, and Björn Schuller, "Speech emotion classification using attention-based LSTM," *IEEE/ACM Transactions on Audio, Speech, and Language Processing, IEEE*, vol. 27, pp. 1675-1685, 2019.
- [96] Özseven, Turgut, "A novel feature selection method for speech emotion recognition," *Applied Acoustics*, vol. 146, pp. 320-326, 2019.
- [97] Bhavan, Anjali, Pankaj Chauhan, and Rajiv Ratn Shah, "Bagged support vector machines for emotion recognition from speech," *Knowledge-Based Systems, Elsevier*, vol. 184, p. 104886, 2019.
- [98] Jino Ancilin, and Alex Milton, "Improved speech emotion recognition with Mel frequency magnitude coefficient," *Applied Acoustics, Elsevier*, vol. 179, p. 108046, 2021.
- [99] Mustaqeem, and Soonil Kwon, "Optimal feature selection based speech emotion recognition using two-stream deep convolutional neural network," *International Journal of Intelligent Systems, Wiley Online Library*, vol. 36, pp. 5116-5135, 2021.
- [100] Wang, Yan and Song, Wei and Tao, Wei and Liotta, Antonio and Yang, Dawei and Li, Xinlei and Gao, Shuyong and Sun, Yixuan and Ge, Weifeng and Zhang, Wei and others,

- "A systematic review on affective computing: Emotion models, databases, and recent advances," *Information Fusion*, vol. 83, pp. 19-52, 2022.
- [101] Burkhardt, Felix and Paeschke, Astrid and Rolfes, Miriam and Sendlmeier, Walter F and Weiss, Benjamin and others, "A database of German emotional speech," *Interspeech*, vol. 5, pp. 1517-1520, 2005.
- [102] Costantini, Giovanni and Iaderola, Iacopo and Paoloni, Andrea and Todisco, Massimiliano, "EMOVO corpus: an Italian emotional speech database," *Proceedings of the ninth international conference on language resources and evaluation (LREC'14)*, pp. 3501-3504, 2014.
- [103] Akçay, Mehmet Berkehan, and Kaya Oğuz, "Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers," *Speech Communication, Elsevier*, vol. 116, pp. 56-76, 2020.
- [104] Ilker Bayram and Ivan W. Selesnick, "Frequency-Domain Design of Overcomplete Rational-Dilation Wavelet Transforms," *IEEE Transactions on Signal Processing*, vol. 57, pp. 2957-2972, 2009.
- [105] Koduru, Anusha and Valiveti, Hima Bindu and Budati, Anil Kumar, "Feature extraction algorithms to improve the speech emotion recognition rate," *International Journal of Speech Technology*, vol. 23, pp. 45-55, 2020.
- [106] Ahirwal, Mitul Kumar and Kose, Mangesh Ramaji, "Audio-visual stimulation based emotion classification by correlated EEG channels," *Health and Technology*, vol. 10, pp. 7-23, 2020.
- [107] Ciolino, Jody D and Diebold, Alicia and Jensen, Jessica K and Rouleau, Gerald W and Koloms, Kimberly K and Tandon, Darius, "Choosing an imbalance metric for covariate-constrained randomization in multiple-arm cluster-randomized trials," *Trials*, vol. 20, pp. 1-10, 2019.
- [108] Sharma, Garima and Umapathy, Kartikeyan and Krishnan, Sridhar, "Trends in audio signal feature extraction methods," *Applied Acoustics*, vol. 158, p. 107020, 2020.
- [109] M. S. Likitha, Srinivasa Reddy Ravuri Gupta, K. Hasitha and A. U. Raju, "Speech based human emotion recognition using MFCC," in *International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)*, 2017.
- [110] Junjie Li, Kunpeng Cheng, Suhan Wang, Fred Morstatter, Robert P. Trevino, Jiliang Tang and Huan Liu, "Feature selection: A data perspective," *ACM computing surveys (CSUR)*, vol. 50, pp. 1-45, 2017.
- [111] Pant, Himanshu and Dhanda, Hitesh Kumar and Taran, Sachin, "Sleep apnea detection using electrocardiogram signal input to FAWT and optimize ensemble classifier," *Measurement*, vol. 189, pp. 110485, 2022.
- [112] Li, Dongdong and Zhou, Yijun and Wang, Zhe and Gao, Daqi, "Exploiting the potentialities of features for speech emotion recognition," *Information Sciences*, vol. 548, pp. 328-343, 2021.
- [113] Sajjad, Muhammad and Kwon, Soonil and others, "Clustering-based speech emotion recognition by incorporating learned features and deep BiLSTM," *IEEE access*, vol. 8, pp. 79861-79875, 2020.

- [114] Xu, Mingke, Fan Zhang, and Wei Zhang, "Head fusion: Improving the accuracy and robustness of speech emotion recognition on the IEMOCAP and RAVDESS dataset," *IEEE Access*, vol. 9, pp. 74539-74549, 2021.
- [115] Issa, Dias, M. Fatih Demirci, and Adnan Yazici, "Speech emotion recognition with deep convolutional neural networks," *Biomedical Signal Processing and Control*, vol. 101894, pp. 59, 2020.
- [116] Liu, Zhen-Tao, Meng-Ting Han, Bao-Han Wu, and Abdul Rehman, "Speech emotion recognition based on convolutional neural network with attention-based bidirectional long short-term memory network and multi-task learning," *Applied Acoustics*, vol. 202, pp. 109178, 2023.
- [117] Giannakopoulos, Theodoros, "A method for silence removal and segmentation of speech signals, implemented in Matlab," *University of Athens, Athens*, vol. 2, pp. 17-23, 2009.
- [118] Lu, Jingyi and Yue, Jikang and Zhu, Lijuan and Li, Gongfa, "Variational mode decomposition denoising combined with improved Bhattacharyya distance," *Measurement*, vol. 151, pp. 107283, 2020.
- [119] Kumar, Avinash and Shah Nawazuddin, S and Ahmad, Waquar, "A Noise Robust Technique for Detecting Vowels in Speech Signals," in *Interspeech*, Interspeech, 2020, pp. 3680-3684.
- [120] Muhammad, Ghulam and Alghathbar, Khaled, "Environment recognition from audio using MPEG-7 features," in *Fourth International Conference on Embedded and Multimedia Computing*, 2009.
- [121] Liu, Zhen-Tao, Min Wu, Wei-Hua Cao, Jun-Wei Mao, Jian-Ping Xu, and Guan-Zheng Tan, "Speech emotion recognition based on feature selection and extreme learning machine decision tree," *Neurocomputing, Elsevier*, vol. 273, pp. 271-280, 2018.
- [122] Khalil, Ruhul Amin, Edward Jones, Mohammad Inayatullah Babar, Tariqullah Jan, Mohammad Haseeb Zafar, and Thamer Alhussain., "Speech emotion recognition using deep learning techniques: A review," *IEEE access*, vol. 7, pp. 117327-117345, 2019.
- [123] Lazarus, Richard S, *Emotion and adaptation*, Oxford University Press, 1991.
- [124] Cannon, Walter B, "The James-Lange theory of emotions: A critical examination and an alternative theory," *The American journal of psychology*, vol. 39, pp. 106-124, 1927.
- [125] Ratcliffe, Matthew, "William James on emotion and intentionality," *International Journal of Philosophical Studies*, vol. 13, pp. 179-202, 2005.
- [126] Damasio, Antonio R, "Emotion in the perspective of an integrated nervous system," *Brain research reviews*, vol. 26, pp. 83-86, 1998.
- [127] Mesquita, Batja and Frijda, Nico H, "Cultural variations in emotions: a review," *Psychological bulletin*, vol. 112, pp. 179-185, 1992.
- [128] Zeng, Zhihong and Pantic, Maja and Roisman, Glenn I and Huang, Thomas S, "A survey of affect recognition methods: audio, visual and spontaneous expressions," in *Proceedings of the 9th international conference on Multimodal interfaces*, 2007.

LIST OF PUBLICATION AND THEIR PROOFS

- [1] Ravi and S. Taran, "A nonlinear feature extraction approach for speech emotion recognition using VMD and TKEO." *Applied Acoustics*, Elsevier, vol. 214, pp. 109667, 2023.
- [2] Ravi and S. Taran, "Emotion Recognition Using Energy Based Adaptive Mode Selection." *Speech Communication*, Elsevier, pp.103228, 2025.
- [3] Ravi and S. Taran , "A novel decomposition-based architecture for multilingual speech emotion recognition." *Neural Computing and Applications*, Springer, vol. 36, pp. 9347-9359, 2024.
- [4] Ravi and S. Taran, " A Filtering Approach for Speech Emotion Recognition Using Wavelet Approximation Coefficient." *Measurement*, Elsevier. (Preparing 2nd revision)
- [5] Ravi and S. Taran, " Multi Level Filltering Approach For Speech Emotion Recognition Using Vmd-Bi-Lstm." *Circuits, Systems & Signal processing*, Springer. (Submitted)
- [6] Ravi and S. Taran, "Emotion Recognition Using Rational Dilation Wavelet Transform for Speech Signal," 2021 7th International Conference on Signal Processing and Communication (ICSC), IEEE, Noida, India, 2021, pp. 156-160.
- [7] Ravi and S. Taran, "Emotion Recognition in Speech Using MFCC and Energy Based Ratio Features," 2024 11th International Conference on Signal Processing and Integrated Networks (SPIN), IEEE, Noida, India, 2024, pp. 367-371.
- [8] Hosain, Mehrab, Anukul Pandey, Sachin Taran, Ravi, Asmar Hafeez, and Nikhil. "Speech emotion recognition using empirical wavelet transform and cubic support vector machine." In *Artificial Intelligence: A tool for effective diagnostics*, pp. 13-1. Bristol, UK: IOP Publishing, 2024.



A nonlinear feature extraction approach for speech emotion recognition using VMD and TKEO

Ravi, Sachin Taran^{*}

Delhi Technological University (DTU), Delhi, India

ARTICLE INFO

Keywords:

Speech signal processing
Emotion recognition
Teager-Kaiser energy operator
Variational mode decomposition

ABSTRACT

Speech emotion recognition (SER) is still a challenging research area in human-computer interaction-based systems. This paper proposed a nonlinear feature extraction technique to improve the classification performance of the SER system. The proposed method explores variational mode decomposition (VMD) with the Teager-Kaiser energy operator (TKEO) for the SER. First, VMD decomposes a speech signal into modes, and then the nonlinear TKEO operator is applied to each mode to obtain a time series. The VMD-TKEO preprocessed signal is used to extract the global features based on Energy, Pitch frequency and Mel frequency cepstral coefficients. The features are statistically examined using the Kruskal-Wallis test. The resultant feature set is examined over the support vector machine and its variants for emotion classification. The Ryerson Audio-Visual database is used for the experiment, and different emotion classification problems are formulated. Finally, the accuracy of the proposed SER architecture is quantitatively analyzed, which outperforms the other existing architectures.

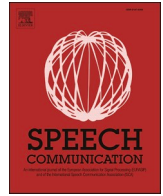
1. Introduction

The advancement of human-computer interaction (HCI) systems owes much to Speech Emotion Recognition (SER). By improving the utility of HCI systems, SER has a wide range of applications in E-learning, robotic interfaces, computer games, entertainment, audio surveillance, clinical studies, and more. For instance, SER can assist educators in comprehending a learner's emotional state and creating a suitable learning environment within the classroom, as stated in [1].

SER is still a difficult task because of the non-stationary behaviour of the speech signal. For example, a human listener can perceive the right emotions of an anonymous speaker only with 60 % recognition accuracy [1]. Further, the limited availability of training data is also a challenging issue. Most SER methods report low accuracy [2–6]. Another challenge in SER is to develop a language-independent system. The accuracy of SER is depended upon language, gender, emotion class, training dataset size, selected feature set, number of emotions, etc. The SER model can be improved by choosing well-designed features that best describe each sense of speech signal [2]. However, the non-stationary behaviour of speech signals requires local and global features to identify emotions. The global features represent statistical aspects like extrema, standard deviation and average value, whereas the local features represent a temporal dynamic. Based on the applications, SER system features are

Teager-Kaiser energy operator (TKEO) based features, speech quality features, Spectral features, and prosodic features [3]. Prosodic features represent the human's perception by analyzing the rhythm and intonation of the speech signal. These features are explored by estimating the speech signal's energy, length and fundamental frequency. At the same time, spectral features deal with the vocal cord characteristics [3]. For the classification of emotions, various classification methods like Recurrent Neural Networks (RNN) [4], Support Vector Machine (SVM) [5], Hidden Markov Model, Neural Networks [6], and Deep learning [7] are used. Researchers from [8–14] employ conventional techniques to classify various combinations of emotions, including two-class classifications examined in [8] and [9]. In [8], a convolutional neural network is trained on the RAVDESS dataset, resulting in a maximum accuracy of 66.41 %. In [9], a bi-directional long short-term memory model is trained on a similar dataset and class, also achieving an accuracy of 70.4 % for the happy class emotion. In [10], the author employs rational dilation wavelet transform decomposition on the same dataset, achieving a maximum recognition accuracy of 83.3 % for a two-class classification. On the other hand, authors in [11–13] use different datasets for four-class emotion classification. In [11], the KNN model was trained using the BERLIN and HINDI datasets, achieving a maximum accuracy of 90 % for the angry class and 70–80 % accuracy for the neutral class. B. Zhang et al. [12] proposed a multi-task learning

^{*} Corresponding author at: Delhi Technological University (DTU), Shahbad Daultpur, New Delhi 110042, India
E-mail address: sachintaran@dtu.ac.in (S. Taran).



Speech emotion recognition using energy based adaptive mode selection

Ravi, Sachin Taran*

Delhi Technological University (DTU), Shahbad Daultpur, New Delhi 110042, India

ARTICLE INFO

Keywords:

Speech signal
Emotion recognition
Variational mode decomposition
Energy estimation

ABSTRACT

In this framework, a speech emotion recognition approach is presented, relying on Variational Mode Decomposition (VMD) and adaptive mode selection utilizing energy information. Instead of directly analyzing speech signals this work is focused on the preprocessing of raw speech signals. Initially, a given speech signal is decomposed using VMD and then the energy of each mode is calculated. Based on energy estimation, the dominant modes are selected for signal reconstruction. VMD combined with energy estimation improves the predictability of the reconstructed speech signal. The improvement in predictability is demonstrated using root mean square and spectral entropy measures. The reconstructed signal is divided into frames, and prosodic and spectral features are then calculated. Following feature extraction, ReliefF algorithm is utilized for the feature optimization. The resultant feature set is utilized to train the fine K- nearest neighbor classifier for emotion identification. The proposed framework was tested on publicly available acted and elicited datasets. For the acted datasets, the proposed framework achieved 93.8 %, 95.8 %, and 93.4 % accuracy on different language-based RAVDESS-speech, Emo-DB, and EMOVO datasets. Furthermore, the proposed method has also proven to be robust across three languages: English, German, and Italian, with language sensitivity as low as 2.4 % compared to existing methods. For the elicited dataset IEMOCAP, the proposed framework achieved the highest accuracy of 83.1 % compared to the existing state of the art.

1. Introduction

In the realm of communication, various methods exist, but speech signals stand out as the swiftest and most innate means for both human-to-human and human-to-machine interaction. Remarkably, humans often possess the ability to discern the emotional nuances of their communication counterparts solely through speech signals. Speech emotion serves as a mechanism for scrutinizing vocal behavior, acting as an indicator for diverse effects such as emotions, moods, and stress, with a specific emphasis on the nonverbal components of speech signal (Hashem et al., 2023). The primary challenge in this domain is the dependency of emotion on language, gender and age. Speech emotion recognition (SER) is becoming a very popular domain of study, especially in the areas of multimedia extraction, human-robot interaction, and human-machine interface. Current SER research focuses on the analytical characteristics of certain acoustic components as well as the investigation of broad qualitative acoustic correlations connected to speech emotions (Quan et al., 2021; Latif et al., 2018; Parlak et al., 2014).

Analyzing an individual's emotional state through speech signals is

still challenging. The process of SER is based on three elements: signal preprocessing, features, and classifiers. There are various methods proposed for the preprocessing of signal. The researchers have proved that SER accuracy is improved after the preprocessing of speech signal with the same feature and dataset. Various traditional and nontraditional approaches have been proposed for speech signal processing (Ravi and Taran, 2024). Traditional approaches include silence removal, voice and unvoiced part detection, noise removal etc. (Weishan Zhang et al., 2017). The commonly used approach shows a higher dependency on language (Soltani et al., 2024; Özseven, Mar. 2019; Ancilin and Milton, 2021; Kwon, 2021; Tuncer et al., 2020). In (Soltani et al., 2024) a method proposed based on Deep Echo State Network and Newman-Watts-Strogatz graph topology. This model tested on three different languages and shows the higher language dependency. The method proposed in (Özseven, Mar. 2019) uses novel feature selection method to reduce computational complexity. The proposed framework is tested for the emotion identification from three publicly available databases, but this framework shows >20 % language sensitivity (Özseven, Mar. 2019). In (Ancilin and Milton, 2021), a language-dependent method using a multiclass support vector machine

* Corresponding author.

E-mail addresses: ravi_2k20phdec12@dtu.ac.in (Ravi), sachintaran@dtu.ac.in (S. Taran).

<https://doi.org/10.1016/j.specom.2025.103228>

Received 30 March 2024; Received in revised form 5 February 2025; Accepted 20 March 2025

Available online 22 March 2025

0167-6393/© 2025 Elsevier B.V. All rights are reserved, including those for text and data mining, AI training, and similar technologies.



A novel decomposition-based architecture for multilingual speech emotion recognition

Ravi¹ · Sachin Taran¹

Received: 23 March 2023 / Accepted: 14 January 2024

© The Author(s), under exclusive licence to Springer-Verlag London Ltd., part of Springer Nature 2024

Abstract

Multilingual speech emotion recognition (MLSER) is a significant and demanding research domain to improve the utility of human–computer interaction systems. Identifying the emotions from the spoken sentence is one of the most challenging tasks due to the dependency of the MLSER system on spoken languages. This study proposes a novel decomposition-based architecture for MLSER. The architecture includes silence removal, mode tuning, signal reconstruction, feature extraction, feature optimization and classification. In preprocessing, the silence part is removed using short-time energy and spectral centroid. After that, variational mode decomposition is applied for signal decomposition, where the improved Bhattacharyya distance is explored for the decomposition mode tuning. The tuned modes are examined for noise removal, and the signal is reconstructed using denoised modes. The spectral and prosodic features are computed from the reconstructed signal. The optimized features are obtained from the extracted features using the ReliefF algorithm. Finally, the fine k-nearest neighbor classifier is explored with optimized features to identify the emotions. For the experiment, three publicly available emotion databases, namely the English language-based Ryerson audio–visual database (RAVDESS), German language-based emotional speech Berlin database (Emo-DB) and Italian emotional speech database (EMOVO), are used. The proposed method yielded 90.7%, 94% and 91.1% accuracy for English, German and Italian language-based database, respectively. A multilingual database is created with these three databases, and the proposed method yields 93.4% accuracy for this database. The proposed framework provides more efficient and minimum language dependency compared to available traditional and deep learning-based approaches.

Keywords Multilingual speech emotion recognition · Speech signal processing · Variational mode decomposition · Feature optimization · Machine learning

1 Introduction

Speech is an effective and easiest way to communicate. Speech is a physiological signal that conveys the speaker's mental state or mode, like neutral, happiness, anger, sadness, etc. [1]. If speech is added with emotion, it becomes more effective for information exchange. It improves the utility of human–computer interaction (HCI) systems in

multiple applications like robot interfaces, computer games, call centers, Web-based E-learning, entertainment, etc. Also, speech is widely used for medical applications to understand a person's physical and mental condition [2].

Speech emotion recognition (SER) provides a deeper understanding of human feelings and helps improve the HCI system's applications. However, the emotion recognition accuracy of the SER system depends on the preprocessing of the raw speech signal, quality of feature selection, classifier selection, the language of the speaker and the training database used. Specifically, the multilingual speech emotion recognition (MLSER) presents a formidable challenge, primarily due to various factors such as variations in signal characteristics, the diverse nature of emotion elicitation methods and the limited availability of data. Numerous efforts by researchers have aimed to

✉ Sachin Taran
sachintaran@dtu.ac.in

Ravi
ravi_2k20phdec12@dtu.ac.in

¹ Department of Electronics and Communication, Delhi Technological University (DTU), Shahbad Daulatpur, New Delhi, Delhi 110042, India



ravi pandey <ravi9887@gmail.com>

Your Submission - Major Revision Required MEAS-D-24-11411R1

2 messages

Measurement <em@editorialmanager.com>
Reply-To: Measurement <support@elsevier.com>
To: Ravi Ravi <ravi9887@gmail.com>

Thu, Apr 10, 2025 at 11:51 AM

Ms. Ref. No.: **MEAS-D-24-11411R1**

Title: A Filtering Approach for Speech Emotion Recognition Using Wavelet Approximation Coefficient Measurement

Dear Mr. Ravi Ravi,

Reviewers have now commented on your paper. You will see that they are advising that you revise your manuscript. If you are prepared to undertake the work required, I would be pleased to reconsider my decision.

For your guidance, reviewers' comments are appended below.

If you decide to revise the work, please submit a list of changes or a rebuttal against each point which is being raised when you submit the revised manuscript. Your cover letter for the revised version should describe how you have incorporated the reviewers' comments, not just stating that you have made changes. In addition, please ensure submit a version that highlights the changes you have made (marked version). The most simple way to do this is to use the "redline" comparison feature in Word to prepare a version that shows all the changes.

Your revised paper is due on the **May 01, 2025**. Manuscripts not resubmitted by **May 01, 2025** will be withdrawn automatically.

To submit a revision, please go to <https://www.editorialmanager.com/meas/> and login as an Author.

Your username is: RRavi-764

If you need to retrieve password details, please go to: [click here to reset your password](#)

NOTE: Upon submitting your revised manuscript, please upload the source files for your article. For additional details regarding acceptable file formats, please refer to the Guide for Authors at: <http://www.elsevier.com/journals/measurement/0263-2241/guide-for-authors>

When submitting your revised paper, we ask that you include the following items:

Manuscript and Figure Source Files (mandatory)

We cannot accommodate PDF manuscript files for production purposes. We also ask that when submitting your revision you follow the journal formatting guidelines. Figures and tables may be embedded within the source file for the submission as long as they are of sufficient resolution for Production. For any figure that cannot be embedded within the source file (such as *.PSD Photoshop files), the original figure needs to be uploaded separately. Refer to the Guide for Authors for additional information.

<http://www.elsevier.com/journals/measurement/0263-2241/guide-for-authors>

Highlights (mandatory)

Highlights consist of a short collection of bullet points that convey the core findings of the article and should be submitted in a separate file in the online submission system. Please use 'Highlights' in the file name and include 3 to 5 bullet points (maximum 85 characters, including spaces, per bullet point). See the following website for more information

<http://www.elsevier.com/highlights>

Graphical Abstract (optional)

Graphical Abstracts should summarize the contents of the article in a concise, pictorial form designed to capture the attention of a wide readership online. Refer to the following website for more information: <http://www.elsevier.com/graphicalabstracts>

On your Main Menu page is a folder entitled "Submissions Needing Revision". You will find your submission record



ravi pandey <ravi9887@gmail.com>

CSSP-D-25-00882 - Submission Confirmation1 message

Circuits, Systems & Signal Processing (CSSP) <em@editorialmanager.com>

Thu, Apr 17, 2025 at 2:46 PM

Reply-To: "Circuits, Systems & Signal Processing (CSSP)" <beno.philomen@springernature.com>

To: "Ravi ." <ravi9887@gmail.com>

Dear Mr .,

Thank you for submitting your manuscript, MULTI LEVEL FILLTERING APPROACH FOR SPEECH EMOTION RECOGNITION USING VMD-Bi-LSTM, to Circuits, Systems, and Signal Processing.

The submission id is: CSSP-D-25-00882

Please refer to this number in any future correspondence.

During the review process, you can keep track of the status of your manuscript by accessing the Journal's website.

Your username is: ravi

If you forgot your password, you can click the 'Send Login Details' link on the EM Login page at <https://www.editorialmanager.com/cssp/>.

Should you require any further assistance please feel free to e-mail the Editorial Office by clicking on "Contact Us" in the menu bar at the top of the screen.

Thank you very much.

With kind regards,
Springer Journals Editorial Office
Circuits, Systems, and Signal Processing

Now that your article will undergo the editorial and peer review process, it is the right time to think about publishing your article as open access. With open access your article will become freely available to anyone worldwide and you will easily comply with open access mandates. Springer's open access offering for this journal is called Open Choice (find more information on www.springer.com/openchoice). Once your article is accepted, you will be offered the option to publish through open access. So you might want to talk to your institution and funder now to see how payment could be organized; for an overview of available open access funding please go to www.springer.com/oafunding. Although for now you don't have to do anything, we would like to let you know about your upcoming options.

This letter contains confidential information, is for your own use, and should not be forwarded to third parties.

Recipients of this email are registered users within the Editorial Manager database for this journal. We will keep your information on file to use in the process of submitting, evaluating and publishing a manuscript. For more information on how we use your personal details please see our privacy policy at <https://www.springernature.com/production-privacy-policy>. If you no longer wish to receive messages from this journal or you have questions regarding database management, please contact the Publication Office at the link below.

In compliance with data protection regulations, you may request that we remove your personal registration details at any time. (Use the following URL: <https://www.editorialmanager.com/cssp/login.asp?a=r>). Please contact the publication office if you have any questions.



JAYPEE INSTITUTE OF INFORMATION TECHNOLOGY, NOIDA

(Declared Deemed to be University under section 3 of UGC Act, 1956)

DEPARTMENT OF ELECTRONICS & COMMUNICATION ENGINEERING

2021 7th INTERNATIONAL CONFERENCE ON SIGNAL PROCESSING AND COMMUNICATION (ICSC)

CERTIFICATE OF PARTICIPATION

This is to certify that **Ravi** of **Delhi Technological University, Delhi, India** has presented a paper entitled **Emotion Recognition Using Rational Dilation Wavelet Transform For Speech Signal** through virtual mode in 2021 7th International Conference on Signal Processing and Communication (ICSC) held on 25th - 27th November, 2021 at Jaypee Institute of Information Technology, Noida, India.

(Dr. Jitendra Mohan)
Organizing Secretary

(Prof. Shweta Srivastava)
General Chair

(Prof. Hari Om Gupta)
General Chair



AMITY UNIVERSITY
UTTAR PRADESH



Department of Electronics & Communication Engineering
Amity School of Engineering & Technology

11th International Conference on Signal Processing and Integrated Networks

Technically Co-Sponsored by



SPIN 2024
21-22 March, 2024

Sponsored by



Certificate of Appreciation

This is to certify that **Mr./Ms./Dr./Prof. Ravi** of **Delhi Technological University, India** has presented his/her paper (online) entitled **Emotion Recognition in Speech Using MFCC and Energy Based Ratio Features** at the **11th International Conference on Signal Processing and Integrated Networks (SPIN 2024)** held on 21-22 March, 2024 in Hybrid Mode at Amity University, Noida, India.

This certificate is awarded for his/her valuable contribution in the success of SPIN 2024.

Dr. Pradeep Kumar
Professor & Deputy Head
Dept of ECE, ASET, AUUP
Organizing Chair (SPIN-2024)

Dr. J. K. Rai
Professor & Head
Dept of ECE, ASET, AUUP
Conference Chair (SPIN-2024)

Dr. Manoj Kumar Pandey
Professor & Director
ASET, AUUP
General Chair (SPIN-2024)



CHAPTER 13

Speech emotion recognition using empirical wavelet transform and cubic support vector machine

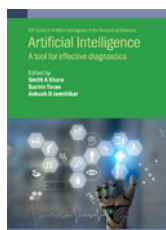
Mehrab Hosain, Anukul Pandey, Sachin Taran, Ravi, Asmar Hafeez and Nikhil

Published November 2024 • Copyright © IOP Publishing Ltd 2024. All rights, including for text and data mining (TDM), artificial intelligence (AI) training, and similar technologies, are reserved.

Pages 13-1 to 13-15

Download complete [PDF book](#), the [ePub book](#) or the [Kindle book](#)

Chapter navigation



Preview

← Previous chapter

Table of contents

Next chapter →

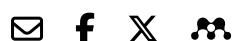
Export citation and abstract

BibTeX

RIS

Permissions

Share this chapter



Abstract

Chapter 13 explores how the human voice is an essential tool for conveying emotions and proposes a novel speech emotion recognition (SER) framework which is based on the combination of empirical wavelet transform and a cubic support vector machine. The system is



DELHI TECHNOLOGICAL UNIVERSITY
 (Formerly Delhi College of Engineering)
 Shahbad Daulatpur, Main Bawana Road, Delhi-42

PLAGIARISM VERIFICATION

Title of the Thesis: **Speech Based Emotion Recognition Using Non-Stationary Decompositions and Optimal Features**

Total Pages: **93**

Name of the Scholar: **Ravi**

Supervisor

(1) **Dr. Sachin Taran**

Department: **Electronics and Communication Engineering**

This is to report that the above thesis was scanned for similarity detection. Process and outcome is given below:

Software used: **Turnitin** Similarity Index: **6%**, Total Word Count: **22234**

Date: _____

Candidate's Signature

Signature of Supervisor

6% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.





Filtered from the Report

- Small Matches (less than 10 words)




Exclusions

- 5 Excluded Sources

Match Groups

-  **107** Not Cited or Quoted 6%
Matches with neither in-text citation nor quotation marks
-  **1** Missing Quotations 0%
Matches that are still very similar to source material
-  **0** Missing Citation 0%
Matches that have quotation marks, but no in-text citation
-  **0** Cited and Quoted 0%
Matches with in-text citation present, but no quotation marks

Top Sources

- 4%  Internet sources
- 3%  Publications
- 2%  Submitted works (Student Papers)

Integrity Flags

0 Integrity Flags for Review

No suspicious text manipulations found.

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.

CURRICULUM VITAE

RAVI

E-mail: ravi9887@gmail.com
ravi9887@rediffmail.com

I wish to be a part of the team that will provide opportunity for accelerated growth and advancements. To enhance and develop my skills and learn about the latest paradigms & processes. Obtain a challenging position in a dynamic environment.

Professional Qualification

Completed high school in 2002 from A.J.I.C Tanda.

Completed **Intermediate** in 2004 from G.I.C Sultanpur.

Completed **B.Tech** (Electronics & Telecommunication) in 2005 from V.I.E.T G.B Nagar.

Completed **M.Tech** (Instrumentation & Signal Processing) in 2015 from V.I.E.T G.B Nagar.

Pursuing Ph.D from DTU, Delhi since 2020.

Journal/Conference/ book Chapter

- [1] Ravi and S. Taran, "A nonlinear feature extraction approach for speech emotion recognition using VMD and TKEO." *Applied Acoustics*, Elsevier, vol. 214, pp. 109667, 2023.
- [2] Ravi and S. Taran , "A novel decomposition-based architecture for multilingual speech emotion recognition." *Neural Computing and Applications*, Springer, vol. 36, pp. 9347-9359, (2024).
- [3] Ravi and S. Taran, "Emotion Recognition Using Energy Based Adaptive Mode Selection." *Speech Communication*, Elsevier, pp.103228, 2025.
- [4] Ravi and S. Taran, " A Filtering Approach for Speech Emotion Recognition Using Wavelet Approximation Coefficient." *Measurement*, Elsevier (1st revision submitted).
- [5] Ravi and S. Taran, "Emotion Recognition Using Rational Dilation Wavelet Transform for Speech Signal," 2021 7th International Conference on Signal Processing and Communication (ICSC), IEEE, Noida, India, 2021, pp. 156-160.
- [6] Ravi and S. Taran, "Emotion Recognition in Speech Using MFCC and Energy Based Ratio Features," 2024 11th International Conference on Signal Processing and Integrated Networks (SPIN), IEEE, Noida, India, 2024, pp. 367-371.
- [7] Ravi, Kumar, Himanshu, Harshul Singh, Indu Banerjee, and Himanshu Mishra. "Design And Power Optimization of ALU Using Adiabatic Logic." In *2024 International*

Conference on Electrical Electronics and Computing Technologies (ICEECT), vol. 1, pp. 1-7. IEEE, 2024.

- [8] Taran, Sachin, Ravi, Smith K. Khare, Varun Bajaj, and G. R. Sinha. "Classification of Alertness and Drowsiness States Using the Complex Wavelet Transform-Based Approach for EEG Records." In *Analysis of Medical Modalities for Improved Diagnosis in Modern Healthcare*, pp. 1-15. CRC Press, 2021.
- [9] Hosain, Mehrab, Anukul Pandey, Sachin Taran, Ravi, Asmar Hafeez, and Nikhil. "Speech emotion recognition using empirical wavelet transform and cubic support vector machine." In *Artificial Intelligence: A tool for effective diagnostics*, pp. 13-1. Bristol, UK: IOP Publishing, 2024.

Personal Strengths

- Self-motivated and a good team player.
- Excellent understanding of Project issues, Ability to work efficiently Under constraints.
- Enjoy learning new areas for self development, procedures & products.

I vouch the authenticity of aforementioned facts.

RAVI