Features for Speech Emotion Recognition

by

S. UTHIRAA
202111065

A Thesis Submitted in Partial Fulfilment of the Requirements for the Degree of

MASTER OF TECHNOLOGY
in
INFORMATION AND COMMUNICATION TECHNOLOGY
to

DHIRUBHAI AMBANI INSTITUTE OF INFORMATION AND COMMUNICATION TECHNOLOGY

May, 2023

Declaration

I hereby declare that

i) the thesis comprises of my original work towards the degree of Master of Technology in Information and Communication Technology at Dhirubhai Ambani Institute of Information and Communication Technology and has not been submitted elsewhere for a degree,

ii) due acknowledgment has been made in the text to all the reference material used.

Uthiraa
21/7/23

S. Uthiraa

Certificate

This is to certify that the thesis work entitled FEATURES FOR SPEECH EMOTION RECOGNITION has been carried out by S. UTHIRAA for the degree of Master of Technology in Information and Communication Technology at *Dhirubhai Ambani Institute of Information and Communication Technology* under my/our supervision.

HPatil
21/07/2023

Prof. Hemant A. Patil
Thesis Supervisor

Acknowledgments

I would like to express my heartfelt gratitude and appreciation to all those who have contributed to the successful completion of my thesis work. Their unwavering support and encouragement have played a pivotal role in my academic journey.

First and foremost, I am deeply indebted to my esteemed guide, Prof. Hemant A.Patil, for his exceptional guidance, mentorship, and expertise. His insightful suggestions, continuous encouragement, and unwavering support have been invaluable throughout the entire thesis process. I am grateful for the opportunities he provided me to grow as a researcher and for his patience and dedication in helping me overcome challenges.

I would like to extend my gratitude to the dedicated members of my Speech Research Lab at DA-IICT. Their camaraderie and willingness to share knowledge have fostered a stimulating research environment, encouraging me to explore new ideas and perspectives.

I am thankful to the staff and authorities of our campus for providing the necessary resources, facilities, and a conducive research environment.Their commitment to promoting academic excellence has been truly commendable.

I would like to acknowledge the support from Ministry of Electronics and Information Technology (Meity), New Delhi, Govt. of India; for their initiative of making National Language Translation Mission (NLTM): BHASHINI, which guided me towards my thesis topic on Emotion Recognition and for their support and funding, which has enabled me to pursue my research aspirations.

Last but not least, I extend my heartfelt appreciation to my family. Their unwavering love, support, and encouragement have been the driving force behind my achievements. Their understanding and belief in my abilities have sustained me throughout this academic endeavor. I am grateful for their sacrifices and for always being there for me.

To all those mentioned above, and to anyone else who has directly or indirectly contributed to the successful completion of my thesis work, I offer my sincere thanks and I am truly grateful for your presence in my academic journey.

Contents

Abstract

The easiest and most effective or natural way of communication is through speech; the emotional aspect of speech leads to effective interpersonal communication. As technological advancements continue to proliferate, the dependence of humans on machines is also increasing, thereby making it imperative to establish efficient methods for Speech Emotion Recognition (SER) to ensure effective human-machine interaction. This thesis focuses on understanding acoustic characteristics of various emotions and their dependence on the culture and language used. It then proposes a new feature set, namely, Constant Q Pitch Coefficients (CQPC) and Constant Q Harmonic Coefficients (CQHC) from Constant Q Transform, which captures high resolution pitch and harmonic information, respectively. Further, this thesis focuses on less explored excitation source-based features and proposes a novel Linear Frequency Residual Cepstral Coefficients (LFRCC) feature set for the same. Phase-based features, namely Modified Group Delay Cepstral Coefficients (MGDCC), is proposed to capture vocal tract and vocal fold information well for emotion classification. The recently developed Automatic Speech Recognition (ASR) model, Whisper, is used to analyze cross-database SER. This thesis extends the LFRCC idea on the infant cry classification problem. Lastly, a local API is developed for SER.

Keywords: *Speech Emotion Recognition, Constant Q Pitch Coefficients, Constant Q Harmonic Coefficients, Linear Frequency Residual Cepstral Coefficients, Modified Group Delay Cepstral Coefficients, Whisper, GMM, CNN, ResNet, TDNN.*

List of Principal Symbols and Acronyms

AESDD	Acted Emotional Speech Dynamic Database
API	Application Programming Interface
ASR	Automatic Speech Recognition
CNN	Convolutional Neural Network
CQCC	Constant Q Cepstral Coefficients
CQHC	Constant Q Harmonic Coefficients
CQPC	Constant Q Pitch Coefficients
CQT	Constant Q Transform
CREMA-D	Crowd Sourced Emotional Multimodal Actors Dataset
DCT	Discrete Cosine Transform
EmoDB	Emotional Database
ESD	Emotional Speech Database
EVC	Emotional Voice Conversion
FFT	Fast Fourier Transform
GDCC	Group Delay Cepstral Coefficients
GFCC	Gammatone Frequency Cepstral Coefficients
GMM	Gaussian Mixture Model
LFCC	Linear Frequency Cepstral Coefficients
LFRCC	Linear Frequency Residual Cepstral Coefficients
MFCC	Mel frequency Cepstral Coefficients

MGDCC	Modified Group Delay Cepstral Coefficients
ResNet	Residual Neural Network
RMSE	Root Mean Square Energy
SER	Speech Emotion Recognition
TDNN	Time Delay Neural Network
TEO	Teager Energy Operator
TESS	Toronto Emotion Speech Set
ZCR	Zero-Crossing Rate

List of Tables

List of Figures

CHAPTER 1
Introduction

Emotion, a term commonly used in everyday life, still lacks a universally accepted scientific definition. Emotion is described as a powerful sensation encompassing feelings such as love or anger or simply referring to feelings in general. It is a mental state triggered by neurophysiological changes that can be linked to thoughts, feelings, behavioral responses, and varying degrees of pleasure or displeasure. Moreover, it represents a complex experience involving consciousness, bodily sensations, and behaviors that reflect the personal significance of something, an event, or a particular situation. Emotion serves as the primary characteristic that distinguishes humans from robots. In the current era of advancing technologies like Artificial Intelligence (AI), our increasing dependence on machines necessitates their meaningful understanding of human emotion. Thus, recognizing emotions is crucial in this regard. Researchers have classified emotion in a *4*-D model, where each dimension—duration, quality, intensity, and pleasure—is independent of the others [22].

The concept of emotion recognition entails the identification of emotions. It is known that speech production is not solely reliant on the vocal tract system, larynx, and lungs. Rather, it involves cognitive processes that transform non-linguistic concepts with communicative intent into linguistically well-formed utterances. This process encompasses linguistic and physiological aspects of speech communication between speakers and listeners [64], [76]. Specifically, it involves the transmission of electrical signals from the brain via motor nerves to activate the larynx and vocal tract system muscles. Consequently, if speech, which conveys emotions, possesses cognitive characteristics, emotion recognition refers to identifying the attributes of emotional states through the observation of ***visual*** and ***auditory*** non-verbal cues. Although emotion can be recognized through speech and facial expressions, this thesis focuses exclusively on analyzing emotions through speech.

As previously mentioned, a substantial gap exists between human and machine-based processing of audio and visual stimuli, preventing machines from accurately identifying a speaker's emotional state. This challenge has led to a new research field called Speech Emotion Recognition (SER) [80]. However, developing an effective emotion recognition model heavily relies on a comprehensive understanding of the acoustics associated with various emotions.

1.1 Problem Statement

This thesis focuses on feature extraction methods for SER. In general, the features for emotion recognition are divided into *4* categories, namely prosodic, excitation source, vocal tract system-related features, and a combination of these [42]. This thesis presents a novel feature extraction method (CQPC and CQHC) that is hypothesized to capture prosodic features. It also touches upon the cultural and linguistic effects on emotions. The thesis also proposes a new feature for the less used excitation source method: LFRCC. The effect of phase-based features on emotions was also studied. Cross-database analysis for emotions was also tested using whisper features.

1.2 Motivation

SER has emerged as a fascinating and relevant area of research in recent years, capturing the attention of researchers, industry professionals, and academia alike. The motivation behind exploring SER for this thesis stems from its profound impact on various applications, such as human-computer interaction, affective computing, and mental health support systems.

Understanding and recognizing human emotions from speech can revolutionize our interactions with technology. Imagine a world where computers can perceive our emotions, adapt their responses accordingly, and provide empathetic support. This would enable more natural and engaging human-computer interactions, enhancing user experiences and opening new avenues for personalized and context-aware technologies.

Furthermore, in affective computing, SER is crucial in building intelligent systems that interpret and respond to human emotions. From virtual assistants and chatbots to automated customer service agents, incorporating SER can enable these systems to detect user emotions and provide appropriate responses. This can enhance customer satisfaction, improve user engagement, and create more meaningful human-like interactions.

Another significant aspect of SER is its application in mental health support systems. Emotions expressed through speech can serve as valuable indicators of an individual's mental well-being. Developing accurate SER models can help in the early detection of mental health conditions, such as depression, anxiety, and stress. By analyzing speech patterns, tone, and prosody, these models can assist mental health professionals in identifying individuals at risk, providing timely interventions, and offering personalized treatment plans.

By conducting research in the field of SER, this thesis aims to contribute to developing robust and accurate emotion recognition systems. Through this research, I hope to pave the way for a future where technology can better understand and respond to our emotions, leading to more empathetic and supportive interactions in various domains.

1.3 Applications and Challenges

SER is a field of study that focuses on the development of algorithms and techniques to automatically detect and analyze emotions from speech signals. SER has numerous applications across various domains and industries. In the field of human-computer interaction, SER can enhance the interaction between humans and machines by enabling devices to understand and respond appropriately to the emotional state of the user. This can lead to the development of more intuitive and empathetic voice assistants, virtual agents, and customer service systems.

In the field of psychology and psychiatry, SER can assist in the diagnosis and treatment of emotional disorders by providing objective measures of emotional states. It can help identify patterns and indicators of different emotions, enabling therapists and researchers to gain insights into patients' emotional well-being and track the effectiveness of therapeutic interventions.

Moreover, SER finds applications in the entertainment industry, particularly in areas such as gaming and virtual reality. By recognizing and adapting to the user's emotional state, game developers can create more immersive and engaging experiences. Additionally, in market research and advertising, SER can be used to evaluate consumer reactions to products, commercials, or brand messaging, providing valuable insights for marketers and advertisers. Some applications of SER are shown in Figure 1.1.

Despite the promising applications, there are several challenges associated with SER. One significant challenge is the subjectivity and variability of emotional expression. Emotions can be influenced by cultural factors, individual dif-

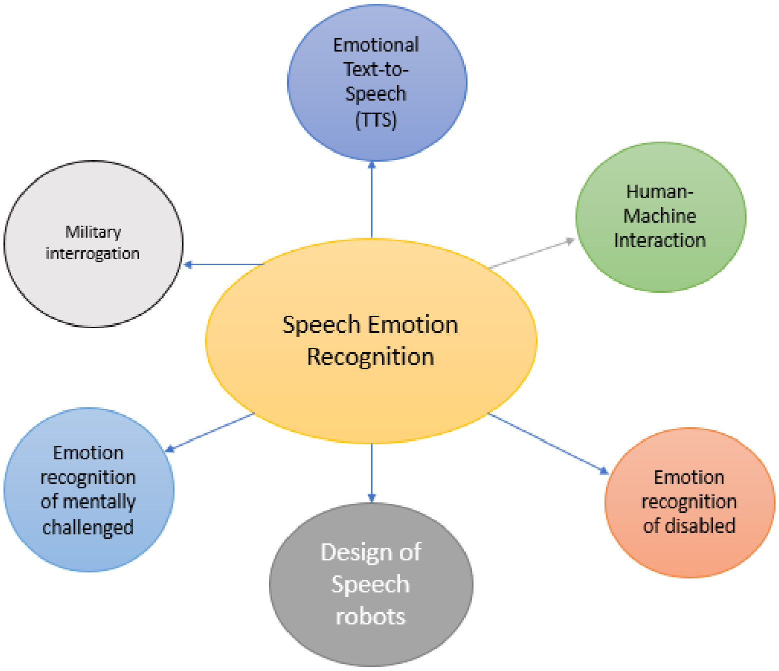

Emotional Text-to-Speech (TTS)
Military interrogation
Human-Machine Interaction
Speech Emotion Recognition
Emotion recognition of mentally challenged
Design of Speech robots
Emotion recognition of disabled

Figure 1.1: Diagram demonstrating significance of emotion recognition.

ferences, and context, making it difficult to develop universal models for emotion recognition. Additionally, the presence of noise, varying speaking styles, and language barriers pose further challenges in accurately detecting and interpreting emotional cues from speech signals.

Another challenge lies in the availability of labeled data for training and evaluating SER models. Collecting large and diverse datasets that encompass a wide range of emotional states can be time-consuming and resource-intensive. Moreover, ensuring the quality and reliability of the labeled data can be challenging, as emotions are subjective and can be interpreted differently by different annotators.

Furthermore, the real-time and online deployment of SER systems introduces additional challenges. Real-time processing requires efficient algorithms and low-latency solutions to provide timely feedback. Handling large volumes of streaming data and ensuring the privacy and security of users' personal information are also important considerations.

1.4 Contributions from this Thesis

- This thesis made an attempt to propose a new prosodic feature set, namely Constant Q Harmonic Coefficients (CQHC) and Constant Q Pitch Coefficients (CQPC) for SER.

- A new feature set, namely Linear Frequency Residual Cepstral Coefficients (LFRCC) was introduced based on less explored excitation source-based features for SER.
- Phase-based Modified Group Delay Cepstral Coefficients (MGDCC) was introduced for SER and its noise robustness qualities were also investigated.
- Recently developed (September 2022) novel ASR model, the whisper is studied. Its features are taken to analyze cross-database performance for SER.
- LFRCC feature set was used for infant cry classification and its performance on the mismatched dataset was the best.
- Local API was built for Assistive Speech Technology, in which the emotion classification model based on CQPC is discussed in detail.

1.5 Organization of the Thesis

- **Chapter 2** presents a detailed study of the previous investigations on SER, databases for emotions, and classifiers used for emotion recognition.
- **Chapter 3** illustrates the speech processing methodologies. Further, this chapter presents the features, classifiers, and performance measures used to evaluate the systems.
- **Chapter 4** kickstarts the work by analyzing emotions and their behavior due to cultural and linguistic differences.
- **Chapter 5** presents the use of timbre-based features (CQHC and CQPC) on SER.
- **Chapter 6** presents the novel LFRCC feature set, based on excitation source information for SER.
- **Chapter 7** presents the novel phase-based MGDCC feature which captures vocal fold and vocal state information for SER. The noise robustness ability of MGDCC is also studied.
- **Chapter 8** proposes a recently developed ASR model, namely, whisper for cross-database analysis in SER.
- **Chapter 9** presents an exploratory work on Infant Cry Analysis using LFRCC.

- **Chapter 10** discusses the details of building a local API for Assistive Speech Technology.

- **Chapter 11** gives the summary, conclusion, limitations, and future research direction on SER.

1.6 Chapter Summary

This chapter gives a brief introduction on emotions and emotion recognition followed by a description of the problem statement. Later in the chapter motivation of this thesis, the application and challenges of the SER are discussed. This chapter is concluded by discussing the contributions and organization of the thesis. In the next chapter, we present the overview of literature search w.r.t. SER.

CHAPTER 2

Literature Survey

2.1 Introduction

The earliest cited work on emotion dates back to 1872, when Darwin in his work "The Expression of the Emotions in Man and Animals" argued that all humans and other animals show emotions through remarkably similar behavior [7]. Furthermore, it was also stated that emotions have an ***evolutionary*** history that could be traced across cultures and species. Following that multiple researchers have worked in this field, some of which include - work by S. Schachter and Singer J in 1962, where they state that the emotional states are a function of physiological arousal and of cognition appropriate to this state [73]. Emotion is studied both by neurobiologists and by emotion theorists; it is important to link psychology and neuroscience to understand emotions better [46]. Culture and society also affect the way people feel and express emotions [72]. All these findings proved why emotion is so complex and so difficult to understand to this date.

2.2 Speech Emotion Recognition (SER)

Initial work on emotion recognition was carried out in late 1999, where Nakatsu R, Tosa N proposed an algorithm for emotion recognition using neural networks. The accuracy obtained was about 50 % [56]. This was then extended by other researchers and now we have multiple emotion recognition features, algorithms, and also datasets are available in various languages [82]. Powerful neural network algorithms are being used to test emotion recognition and accuracy rates have increased ever since [80]. Cognitive features in emotion recognition was also analyzed side-by-side by researchers as their correlation with emotions was found long back [77], [89]. Figure 2.1 shows the detailed structure for SER.

For emotion recognition, the first step needed is to understand its database. There are three types of emotion databases, namely *acted*, *elicited*, and *simulated*

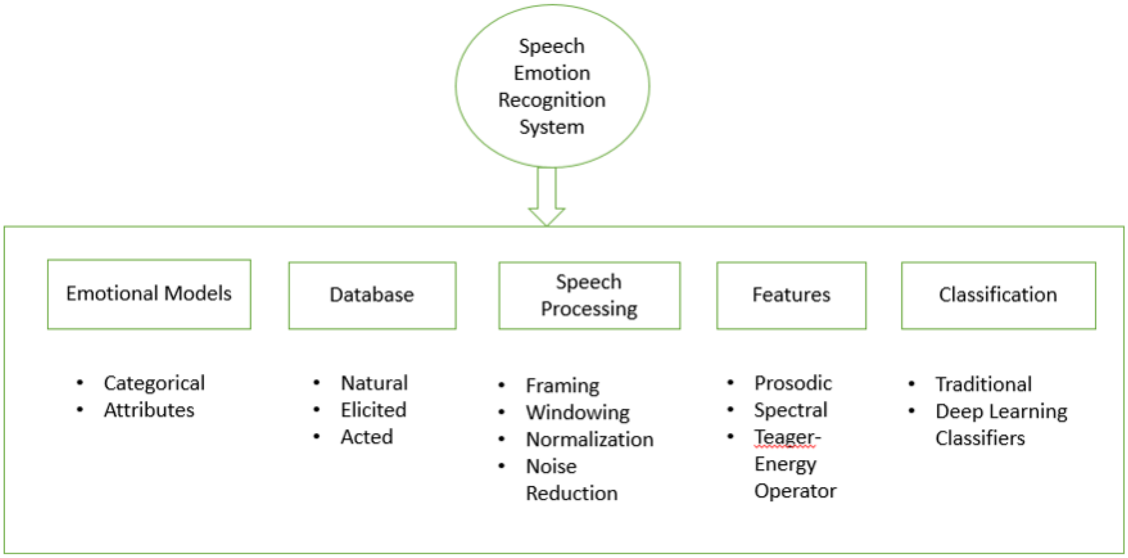


Figure 2.1: Speech Emotion Recognition.

emotions [42], [15]. ***Acted*** emotions refer to emotions that are deliberately portrayed or acted out by individuals, ***Elicited*** emotions are emotions that are intentionally triggered or evoked in individuals through various means. This can be done through external stimuli, such as emotional pictures, videos, or stories, or through interpersonal interactions or specific situations designed to elicit certain emotional responses and ***Natural*** emotions refer to genuine, spontaneous emotional experiences that occur in everyday life situations without any deliberate manipulation or elicitation [80]. The database used in this thesis is of *acted* emotions.

After the database, it is important to understand the features used for SER. Various features are employed for this purpose [26], but the major *four* categories are developed, namely, prosodic, excitation source-based, vocal tract-based, and a combination of the aforementioned features. Prosodic features refer to the suprasegmental aspects of speech that go beyond individual phonemes or words. They involve the rhythm, stress, intonation, and pitch patterns used in spoken language. Prosody plays a crucial role in conveying meaning, emphasis, and emotional expression in communication. Excitation source features (like LP residual) refer to acoustic properties that capture characteristics of the vocal source or laryngeal activity during speech production. These features can provide valuable information for SER; It quantifies the presence or absence of voicing in speech. Emotions may be associated with variations in voicing strength due to differences in vocal fold behavior, these are points in the speech signal that mark the instant of glottal closure during each vocal fold cycle. The analysis of Glottal Closure Instants (GCIs) can reveal timing characteristics related to emotions, etc. Vocal tract features (like MFCC, GFCC) capture the acoustic properties related to the resonant characteristics of the vocal tract during speech production. These features provide valuable insights into the articulatory aspects of emotions expressed through speech. One crucial vocal tract feature used in SER is Formants. Formants represent the resonant frequencies of the vocal tract during speech. They are closely associated with vowel sounds and can provide information about the shape and size of the vocal tract. Emotional states may be reflected in formant patterns, such as variations in their frequency and bandwidth, which can be indicative of different articulatory configurations and emotional expressions.

After extracting emotion features, it is sent to classifiers for testing [40]. Traditional machine learning models, such as Gaussian Mixture Model (GMM), Hidden Markov Model (HMM), K Nearest Neighbour (KNN), Bayes classifier, Support Vector Machine (SVM), random forest, etc were used initially. With the development of Artificial Intelligence (AI) and the emergence of Deep Learning (DL), SER classifications were also shifted to deep learning models. Their ability to automatically learn complex features, handle large amounts of data, model nonlinear relationships, and adapt to diverse tasks and domains make deep learning an increasingly popular and powerful approach in the field of SER. The most commonly used DL models are, Convolutional Neural network (CNN), Recurrent Neural network (RNN), Time Delay Neural Network (TDNN), Long Short Term Memory (LSTM), Residual Neural Network (ResNet), etc.

The previous research primarily concentrated on the lower frequency segments of the speech signal in the context of SER [78]. This served as a driving force behind the creation of CQPC, derived from CQT, as a technique for SER. Furthermore, the limited utilization of the excitation source and phase-based characteristics inspired the adoption of LFRCC and MGDCC, respectively, for SER.

2.3 Chapter Summary

In this chapter, we discussed the attempts made for understanding emotions and the development of SER. The database, features, and classifiers used for SER are studied in detail. In the next chapter, we discuss the experimental setup used for this thesis.

CHAPTER 3

Experimental Setup

Once the problem statement is comprehended, conducting experiments becomes imperative in order to test the formulated hypothesis and address the issue at hand.

3.1 Introduction

This chapter provides a comprehensive overview of the experimental setup employed for SER. It encompasses the description of fundamental speech processing methodologies, such as pre-processing, feature extraction, and post-processing. Additionally, it outlines various datasets and classifiers utilized for SER task. The state-of-the-art features used for performance comparison with the proposed features are also presented. Finally, the chapter discusses the various performance evaluation measures employed to assess the effectiveness of the model.

3.2 Basic Speech Processing Methodologies

3.2.1 Pre-Processing

The primary stage of speech signal processing is pre-processing, which serves to transform the raw input speech into a format that is better suited for extracting features. This involves distinguishing between voiced and unvoiced regions, eliminating background noise, etc. by employing various techniques such as pre-emphasis, framing, windowing, normalization, noise reduction, and silence removal [57]. In this study, framing, windowing, and normalization functions were utilized to prepare the raw input for effective feature extraction.

Framing

Speech framing, also referred to as speech segmentation, is a crucial process in overcoming several challenges related to SER. During natural speech production, emotions tend to fluctuate due to the non-stationary nature of the speech signals. However, the speech itself remains consistent, even within very short duration, typically ranging from 20 to 30 milliseconds. By dividing the speech signal into fixed-length segments, known as speech frames, it becomes possible to estimate semi-fixed and local features. To maintain the connection and information between these frames, intentionally overlapping them by 30% to 40% is beneficial. Consequently, utilizing fixed-size frames becomes suitable for classifiers such as Artificial Neural Networks (ANNs), while retaining the emotional information contained within the speech [91].

Windowing

After the speech signal is blocked into frames, the next step involves applying a window function to each frame. When performing Fast Fourier Transform (FFT) on the signal, discontinuities at the edges can cause spectral leakage, which can be mitigated by employing a windowing function [16]. One commonly used type of windowing function is the Hamming window, defined below -

$$w(n) = 0.54 - 0.46 \cos\left(2\pi n / M - 1\right), \tag{3.1}$$

where w(n) is the window function to extract speech frame, and window size is M, where 0 ≤ n ≤ M - 1 [91].

Normalization

Normalization is a technique employed to adjust the sound volume to a standardized level. In this process, the signal sequence is divided by the highest value of the signal, ensuring that each sentence maintains a comparable volume level [57]. It is calculated using the formula given-

$$z = (x - \mu/\sigma), \tag{3.2}$$

where μ is the mean, and σ is the standard diviation of speech signal [91].

3.2.2 Feature Extraction

It plays a crucial role in speech signal processing as it helps capture relevant information from speech signal for further analysis and classification. As discussed in Chapter 2, various features, such as prosodic, vocal tract, and excitation source-based features, are utilized for SER to effectively capture emotion-specific information. Additionally, depending on the requirements, local and/or global features can be extracted. The resulting numerical values from the feature extraction process are organized into a vector known as the *feature vector* in the pattern recognition and machine learning field. This vectorization aids in reducing the overall signal size, allowing for a more focused and efficient implementation by prioritizing the relevant portions of the signal.

3.2.3 Post-Processing

Post-processing operations play a vital role in enhancing classifier efficiency by mapping the feature space to another space. These operations encompass tasks such as normalization, reducing dimensions of feature vector, incorporating velocity and acceleration coefficients, among others. By carrying out these operations, distortions in the processed input speech are eliminated, ensuring that the resulting data is suitable for further classification by the classifier.

3.3 Details of Dataset Used for SER

In this thesis *five* emotional speech corpora are utilized for testing the proposed features for SER task.

3.3.1 Acted Emotional Speech Dynamic Database (AESDD)

The aforementioned corpus, which was created in 2018, is a freely accessible *greek* dataset designed for SER [58]. It encompasses 500 instances of ***acted*** emotional speech, featuring five distinct emotions: anger, happiness, disgust, fear, and sadness. The dataset consists of recordings from five actors, including three females and two males, with each actor providing 20 utterances per emotion [5]. This corpus was used to analyse emotions using spectrograms, energy and with TEO profile of speech signal.

3.3.2 Emotional Speech Database (ESD)

The recently developed ESD corpus [94], established in 2021, comprises 350 parallel utterances delivered by 10 native English speakers (5 males and 5 females) and 10 native Mandarin speakers (5 males and 5 females). This corpus encompasses emotions, such as anger, happiness, neutrality, sadness, and surprise, with audio samples recorded at a sampling rate of 16 kHz. The selection of this dataset was motivated by its relatively large-scale size, featuring multiple speakers, and its availability to the public, coupled with favorable recording conditions [95]. This dataset was used to examine the influence of *cultural* and *linguistic* disparities between English and Mandarin language on emotions and to test the effectiveness of proposed CQHC, CQPC, and whisper features for SER.

3.3.3 Toronto Emotion Speech Set (TESS)

In this dataset [63], a a collection of 200 specific words that were uttered within the introductory phrase "Say the word–". Two actresses, aged 26 and 64 respectively, were involved in the recordings, which captured each word being expressed with seven different emotions: anger, disgust, fear, happiness, pleasant surprise, sadness, and a neutral tone. There are 400 *.wav* files for each emotion, so in total, 2800 files out of which this work used *five* emotions, namely, anger, happiness, sadness, surprise, and neutral (i.e., 2000 files) for 5 emotions. This dataset was used to analyse the performance of CQHC, CQPC, and whisper features for SER.

3.3.4 German Emotional Database (EmoDB)

The EmoDB dataset is a collection of German speech samples uttered by ten actors, consisting of five male and five female actors. Under favorable recording conditions, the actors recorded ten German phrases expressing seven emotions: anger, joy, neutral, sadness, disgust, boredom, and fear [21]. This work used *four* emotions, namely, anger, joy, neutral, and sadness (339 *.wav* files) for performance analysis of LFRCC and MGDCC on emotions.

3.3.5 Crowd Sourced Emotional Multimodal Actors Dataset (CREMA-D)

The CREMA-D dataset [23], contains 7442 acted utterances from 91 actors. The dataset encompasses six emotions, with the intensity of emotion specified as high, medium, low, or unspecified, featuring 48 male and 43 female actors. This work used 896 utterances per emotion (767 for neutral) for training, while 124 utterances per emotion were used for testing (105 for neutral), and 251 utterances per emotion were used for validation (215 for neutral) for whisper feature analysis. However, it should be noted that the neutral emotion category contains fewer utterances than the other emotions in the dataset.

3.4 State-of-the-art Features for SER task

Some of the widely used features for SER are explained below. These features were used as the baseline to analyse the performance of proposed features.

3.4.1 Mel Frequency Cepstral Coefficients (MFCC)

Mel Frequency Cepstral Coefficients (MFCCs) are widely used to extract essential features in speech and audio signal processing. These coefficients are derived from the Mel-scale, a perceptual pitch scale closely resembling the human auditory system's frequency response [79]. If the input frequency in Hertz is considered as f, then the corresponding frequency (m) in Mel scale is given by [35]:

$$m = 2595 * log(1 + f/700) \tag{3.3}$$

The process of extracting MFCCs involves several key steps (Figure 3.1):

- Pre-emphasis: The input speech signal is typically pre-emphasized to amplify higher frequencies and compensate for spectral density roll-off using a high pass filter.

- Framing: The pre-emphasized speech signal is divided into short overlapping frames, usually ranging from 20 to 40 ms. Overlapping frames ensure information continuity between adjacent frames and thus, avoids the abrupt chopping of speech information.

- Windowing: Each frame of the speech signal is multiplied by a window function, often a Hamming window, to alleviate spectral leakage caused by abrupt frame blocking.

- Fast Fourier Transform (FFT): The windowed frames undergo Fourier transform operation to obtain the frequency spectrum. The power spectrum is

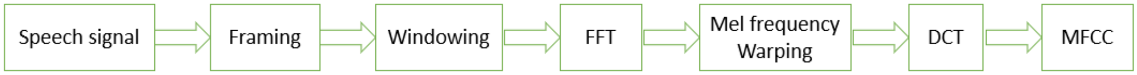


Figure 3.1: MFCC Extraction.

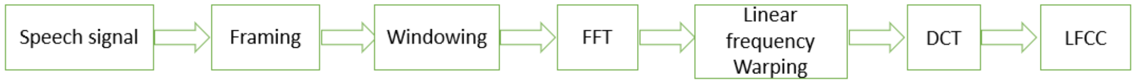


Figure 3.2: LFCC Extraction.

computed by squaring the magnitude of the complex Fourier transform coefficients.

- Logarithm: The logarithm of the filterbank energies is taken to transform the linear-scale magnitudes into a logarithmic scale. This transformation imitates the non-linear loudness perception of the human ear.

- Discrete Cosine Transform (DCT): Finally, a discrete cosine transform is applied to the logarithmic filterbank energies. The resulting coefficients, MFCCs, effectively capture the speech signal's spectral characteristics. Further DCT is applied for feature vector dimentionality reduction, feature vector decorrelation, and energy compaction.

3.4.2 Gammatone Frequency Cepstral Coefficients (GFCC)

Gammatone Frequency Cepstral Coefficients (GFCCs) are a modified version of the well-known MFCC feature extraction. While MFCC employ Mel subband filters for feature extraction in speech and audio signal processing, GFCC utilize gammatone filters as an alternative. They show robustness against noise and acoustic change [49] thereby, making it better than MFCC, thus used for comparison with proposed features.

3.4.3 Linear Frequency Cepstral Coefficients (LFCC)

LFCC captures the spectral characteristics of the speech signal, similar to MFCC, however use linearly-spaced subband filters instead of the Mel filters. Steps to extract LFCC is shown in Figure 3.2.

3.5 Details of Classifier Used

Multiple classifiers were employed to assess the impacts of the proposed features, and their specifics are provided in the following Sub-sections.

3.5.1 Convolutional Neural Network (CNN)

Convolutional Neural Networks (CNN) [9] are deep learning models that leverage the convolution operation to process data within their architecture. This convolution occurs between multidimensional input data and multidimensional filter weights, referred to as *kernels*. Following the convolutional operation, there are subsequent pooling layers and non-linear activation operations. Together, these three operations form a convolutional layer, responsible for extracting features from the input data. CNN models also incorporate fully-connected layers of perceptrons for classification purposes.

Convolution Operation

Within the framework of CNN, the convolution operations involve sliding the kernel across the input matrix and processing the data accordingly. It is worth noting that the kernel size is typically smaller than the input matrix. The convolution operator is mathematically represented as:

$$\text{Output}[i,j] = \sum_{m=0}^{\text{kernel\_height}-1} \sum_{n=0}^{\text{kernel\_width}-1} \text{Input}[i+m, j+n] \cdot \text{Kernel}[m,n] + \text{Bias}, \tag{3.4}$$

where "Output" represents the output feature map, "Input" denotes the input feature map, "Kernel" refers to the kernel weights, and "Bias" represents the bias term. The summations iterate over the *height* and *width* of the kernel, and the resulting sum is computed at each spatial location (*i,j*) in the output feature map.

This operation has a profound impact on the network's ability to extract and learn meaningful features from input data. By convolving filters (kernels) across the input, CNNs are able to detect various patterns, edges, and textures at different spatial scales. This enables the network to build a hierarchical representation of features, with lower layers capturing simple features and higher layers capturing more complex ones. The parameter sharing mechanism in convolutional layers reduces the number of learnable parameters, promoting efficient learning and better generalization. Additionally, the translation invariance property of convolutions allows CNNs to recognize features regardless of their exact location in the input, contributing to robustness in tasks such as object recognition.

Padding Operation

Padding plays a critical role in CNN by adding extra rows and columns of zeros to input feature maps. This operation serves multiple purposes. Firstly, it maintains the spatial dimensions of feature maps throughout the convolutional layers, preserving vital spatial information, especially in deeper layers. Secondly, it enables better control over the output size after convolution, ensuring compatibility with subsequent layers and facilitating architectural alignment. Moreover, padding ensures comprehensive coverage of the feature map borders during convolution, facilitating the capture of significant edge and boundary details. Lastly, it mitigates the border effect by preventing a decrease in spatial resolution near the feature map edges, leading to more accurate representations. It is calculated using the formula given below.

$$\text{Padding Size} = \left\lceil \frac{(\text{Output Size} - 1) \times \text{Stride} - \text{Input Size} + \text{Kernel Size}}{2} \right\rceil \tag{3.5}$$

Stride Convolution

Stride convolution in CNN involves moving the convolutional kernel across the input feature map with a specified step size, known as the *stride*. By adjusting the stride value, CNN can control the downsampling of the spatial dimensions of the feature map. A larger stride reduces the spatial resolution, allowing for faster processing and decreased computational complexity. However, it may lead to information loss and less precise feature localization. On the other hand, a smaller stride preserves spatial detail but increases computational requirements. Stride convolution is often used in conjunction with padding to control the output size and maintain compatibility with subsequent layers. It enables CNN to strike a balance between computational efficiency and feature representation, depending on the specific needs of the task. It is calculated using:

$$\text{Stride} = \frac{\text{Input Size} - \text{Kernel Size}}{\text{Output Size} - 1}. \tag{3.6}$$

It is important to note that the stride should be a positive integer greater than or equal to 1. If the calculated stride is a non-integer value, it is often rounded up to the nearest whole number to ensure a valid stride.

Activation Layers

The activation function is a vital component that introduces non-linearity to the network. It applies a non-linear transformation element-wise to the output of the convolutional layer or other preceding layers. The activation function plays a crucial role in allowing the network to learn complex patterns and make more expressive predictions. Popular activation functions include ReLU, which sets negative values to zero and keeps positive values unchanged, and sigmoid, which squashes the output between 0 and 1. Other functions such as tanh, leaky ReLU, and maxout provide different characteristics and flexibility in capturing non-linear relationships. The activation function enables CNN to model complex features and make accurate classifications, contributing to their effectiveness in computer vision tasks and deep learning applications.

Pooling Layers

Pooling layers are an essential part of CNN that aid in reducing the spatial dimensions of feature maps while preserving important information. These layers divide the feature maps into non-overlapping regions and apply operations like max pooling or average pooling to extract the most significant values within each region. The main purposes of pooling layers are to decrease computational complexity, create translation invariance, and extract salient features. By reducing the dimensions, pooling layers make the network more efficient and capable of capturing important patterns while discarding redundant details.

Architecture Details

- **LFRCC on infant cry analysis:** CNN was trained using the sigmoid activation function. It employes 5 convolutional layers, each with output 16, 64, 64, 16, and 16, respectively, with kernel size of 3×3. Each convolutional layer was followed by max-pooling of 2 and ReLU function. 2 fully-connected layers were employed with a dropout of 0.25.

- **CQPC and CQHC on SER:** The model is trained using stratified *10*-folds cross-validation strategy with train and validation split of *80* % and *20* % using *adam* optimizer, *categorical cross-entropy* as a loss function, *accuracy* as the evaluation metric, and (5x5) as the kernel size. The learning rate used is *0.001* and batch size is taken as *32*.

3.5.2 Residual Neural Network (ResNet)

ResNet, also known as Residual Network, is a deep convolutional neural network architecture proposed by Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun in 2016 [32]. Its primary objective is to overcome the vanishing gradient problem when training deep neural networks.

The fundamental concept behind ResNet is the utilization of residual connections, also called skip connections or shortcut connections. These connections enable the network to bypass specific layers and directly transmit information from one layer to another. This technique effectively mitigates the vanishing gradient problem and facilitates the training of significantly deeper networks.

In a conventional neural network, each layer learns to map its input to its output. However, in a ResNet, each layer learns a residual mapping, which represents the difference between its input and output. The layer's output is obtained by adding the residual mapping to the input, as illustrated by the equation:

$$y = F(x) + x. \tag{3.7}$$

Here, y represents the layer's output, F(x) represents the learned residual mapping, and x represents the layer's input. By explicitly learning the residuals, ResNet enables the network to focus on learning residual functions rather than attempting to learn the entire mapping from scratch. This characteristic significantly simplifies the training of deep networks.

The ResNet architecture is composed of multiple residual blocks that are stacked sequentially. Each residual block typically incorporates two or three convolutional layers, accompanied by batch normalization and activation functions such as ReLU. The skip connections in ResNet can be implemented as identity mappings, directly passing the input to the output, or using 1x1 convolutional layers to adjust the dimensions between input and output.

ResNet has proven to be a highly effective architecture in various computer vision tasks, such as image classification, object detection, semantic segmentation, and image captioning. Its utilization of residual connections has revolutionized deep network training by facilitating gradient flow and enabling the successful training of deep neural networks with superior performance.

Architecture Details

- **LFRCC in SER:** The ResNet model used for this work comprises of 3 residual blocks of size (64,32,2), (32,16,2), and (16,16,2) with a kernel size of 3, strides of *2*, average pooling, loss function as *categorical cross-entropy*, learning rate of *0.001*, and *Stochastic Gradient Descent* as optimiser and is implemented in pytorch.

3.5.3 Time Delay Neural Network (TDNN)

The Time Delay Neural Network (TDNN) is an architecture specifically designed for processing sequential data, such as speech and audio signals, time series data, and temporal patterns. It was originally proposed by Alex Waibel and colleagues in 1989 [45].

The key feature of the TDNN architecture is its capability to capture temporal dependencies by incorporating time delays within the network structure. Unlike conventional feedforward neural networks that only consider the current input, TDNN takes into account information from multiple time steps.

In a TDNN, each neuron in a layer is connected to neurons in the previous layer at different time delays. These time delays enable the network to incorporate a window of past information in addition to the current input. By considering the temporal context, TDNN can effectively capture evolving patterns and relationships over time.

Typically, TDNN consists of multiple layers with shared weights across different time delays. This weight sharing allows the network to generalize across various time steps and extract meaningful features from the input sequence. The output of each layer is computed by convolving the inputs with the corresponding weights, applying a non-linear activation function, and combining the results.

A popular variant of TDNN is the Time-Delay Neural Network for Speech Recognition (TDNN-F), which has gained significant usage in speech recognition systems. TDNN-F extends the basic TDNN by incorporating additional layers, such as fully connected layers and softmax layers, to enable classification tasks on sequential data.

TDNN has demonstrated effectiveness in diverse applications involving sequential data, including speech and speaker recognition, natural language processing, and music analysis. Its ability to model temporal dependencies and capture long-range patterns makes it particularly suitable for tasks, where the ordering of inputs is critical.

Architecture Details

- **LFRCC in SER:** In TDNN [62], the layer-wise input and output dimensions for each layer of the TDNN were (39, 64), (64, 128), (128, 128), (128, 64), and (64, 64). Dropout at *0.2*, attention pooling, loss function as *categorical cross-entropy*, a learning rate of *0.001*, and *Adam* optimizer are used for classification.

3.5.4 Gaussian Mixture Model (GMM)

The Gaussian Mixture Model (GMM) is a statistical framework employed to represent the probability distribution of a dataset [70]. It assumes that the dataset originates from a combination of multiple Gaussian distributions, referred to as components. The GMM aims to approximate the underlying data distribution by learning the parameters associated with these components.

The GMM represents the data as a weighted sum of Gaussian distributions. Each Gaussian distribution represents a distinct cluster within the data, and the model assigns weights to these components to indicate their respective contributions to the overall distribution. The parameters of the GMM encompass the means and variances of the Gaussian components and the weights assigned to each component.

These parameters are estimated using the Expectation Maximization (EM) algorithm, which iteratively maximizes the likelihood of the data given the current parameter estimates. Once the model is trained, it becomes capable of generating new data points by sampling from the learned distribution.

Architecture Details

512 number of Gaussian mixtures were used for training the Baby Chilanto dataset. However, a 128 number of Gaussian mixtures were used to train the in-house (i.e., DA–IICT) dataset as there are less number of data samples. The scores were computed using the log-likelihood function.

3.6 Performance Evaluation Measures

The performance of classifer models and feature sets were computed using multiple performance measures. The details of the same are mentioned in the following sub-Sections.

3.6.1 k-Fold Cross Validation

k-fold cross-validation is a widely used technique in machine learning and statistics to evaluate how well a model performs and generalizes. It is especially beneficial when data is limited and allows us to estimate the model's performance on unseen data.

k-fold cross-validation operation:

- The original dataset is divided into k subsets, each having roughly the same amount of data. These subsets, known as "folds," are created.
- The model is trained and assessed k times, each iteration using a different fold as the validation set while utilizing the remaining k-1 folds as the training set.
- In each iteration, the model is trained on the training set and then evaluated on the validation set using a selected evaluation metric, such as accuracy or mean squared error.
- The performance scores obtained from each iteration are averaged together, resulting in a single performance metric representing the model's overall performance.

One of the key advantages of k-fold cross-validation is its ability to provide a more dependable estimate of the model's performance when compared to a single train-test split. Repeating the process k times and averaging the results reduces the influence of a specific train-test split on the performance evaluation, resulting in a more robust assessment of the model's capabilities.

3.6.2 Confusion Matrix

A confusion matrix is a commonly employed tabular tool for assessing the effectiveness of a classification model. It presents a concise overview of the model's predictions on a specific test dataset, allowing for a comparison against the true labels of the data. The matrix comprises four essential elements:

- **True Positive (TP)**: The number of samples that were correctly predicted as positive.
- **True Negative (TN):** The number of samples that were correctly predicted as negative.

- **False Positive (FP):** The number of samples that were incorrectly predicted as positive.
- **False Negative (FN):** The number of samples that were incorrectly predicted as negative.

These metrics aid in evaluating the model's ability to correctly identify positive and negative instances, providing valuable insights into areas that could be enhanced.

3.6.3 % Classification Accuracy

Accuracy is the most simplified and powerful performance metric employed to evaluate the effectiveness of traditional machine learning and deep learning models. The accuracy (in %) is defined as:

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}} \times 100\%. \tag{3.8}$$

Classification accuracy is frequently employed as an evaluation metric, particularly in datasets with balanced classes, where the number of instances in each class is similar. However, in the case of imbalanced datasets, where one class prevails, accuracy may not be the most suitable metric. In such scenarios, alternative metrics, such as precision, recall, and F1-score can offer a more comprehensive evaluation of the model's performance.

3.6.4 $F1$-Score

The F1-score [30] is a widely utilized metric in classification tasks for evaluating the performance of a model. By combining precision and recall, it offers a well-rounded measure of the model's accuracy.

$$F1\text{-score} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}, \tag{3.9}$$

where precision captures the ratio of accurately predicted positive instances among all instances predicted as positive. On the other hand, recall measures the proportion of correctly predicted positive instances out of all actual positive instances.

The F1-score, ranging from 0 to 1, serves as an indicator of performance. A score of 1 denotes ideal precision and recall, while a score of 0 signifies the poorest performance.

In scenarios involving imbalanced datasets or when both precision and recall hold equal importance, the F1-score proves invaluable.

3.6.5 Jaccard Index

The Jaccard Index, which is alternatively referred to as the Jaccard similarity coefficient or Jaccard coefficient, serves as a metric for quantifying the similarity between two sets [17]. It provides a measure of how much the sets overlap or share common elements.

3.6.6 Mathews' Correlation Coefficient (MCC)

It shows the degree of association between the expected and actual class [52].

$$\text{MCC} = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}. \tag{3.10}$$

The range of Matthews' Correlation Coefficient (MCC) extends from -1 to +1, with a score of +1 denoting a flawless prediction, 0 representing a random prediction, and -1 indicating a completely contradictory prediction.

MCC proves particularly valuable when handling imbalanced datasets or situations where the classes possess varying sizes. It takes into account both true and false positives as well as true and false negatives, offering a balanced evaluation of the classifier's performance that remains unaffected by class imbalance.

3.6.7 Hamming Loss

Hamming Loss serves as a metric to evaluate the accuracy of multi-label classification models. It quantifies the proportion of incorrectly predicted labels in relation to the total number of labels across all instances [25].

$$HammingLoss = \frac{1}{N \cdot L} \sum_{i=1}^{N} \sum_{j=1}^{L} (y_{ij} \oplus \hat{y}_{ij}), \tag{3.11}$$

where, N corresponds to the total number of instances. L represents the total number of labels, $y_i j$ signifies the ground truth label for instance i and label j, where it is assigned a value of 1 if the label is present, and 0 otherwise. The symbol $\hat{y}_i j$ denotes the predicted label for instance i and label j, with a value of 1 if the label is predicted, and 0 otherwise. The operator $\oplus$ denotes the XOR (exclusive OR) operation.

3.7 Chapter Summary

This chapter discusses the experimental setup needed for SER. The necessary steps involved in the detection i.e., the speech processing methods were briefly explained. Furthermore, the chapter delves into a comprehensive description of the datasets employed in the thesis experiments. It proceeds by elucidating the cutting-edge features utilized for SER in this study, and subsequently presents a thorough account of the classifiers employed along with their corresponding performance evaluation metrics.

CHAPTER 4
Analysis of Emotions

4.1 Introduction

To develop an efficient SER, it is essential to understand the acoustics of various emotions in speech signal. In particular, humans can remarkably convey and perceive emotions in speech using various prosodic cues, such as loudness (amplitude), pitch, duration of speech sound units, intonation (i.e., derivative of pitch contour), and energy of speech wave. In this chapter, we analyze five emotions from the dataset, namely, anger, fear, disgust, happiness, and sadness. We use narrowband spectrograms, their energies, and their Teager Energy Operator (TEO) profile to differentiate these emotions. Then we study the impact of culture and language on emotions.

4.2 Narrowband Spectrograms

Spectrogram is a visual representation of spectral enenrgy density (z-axis) for an acoustic signal, which is a function of time (X- axis), and frequency (Y-axis). Energy distribution and pauses are also measured from this plot depending on the size of analysis window (a function used for Fourier-based spectral analysis), different levels of frequency/time resolution is achieved, i.e., a long window (atleast 2 pitch periods) gives good frequency resolution as harmonic lines (seen as horizontal striations in the time-frequency domain) are resolved, this is also known as ***narrowband*** spectrogram. Similarly, a short window (duration less than a single pitch period) shows the formants of the vocal tract with individual pitch periods appearing as vertical lines (striations), called ***wideband*** spectrogram. In this work, we obtained the narrowband spectrograms for male and female speakers for 5 emotions using the wavesurfer tool. Figure 4.1 and Figure 4.2 depicts the spectrograms obtained for a sentence uttered by a male and a female speaker, respectively. Narrowband spectrograms are employed rather than their wideband

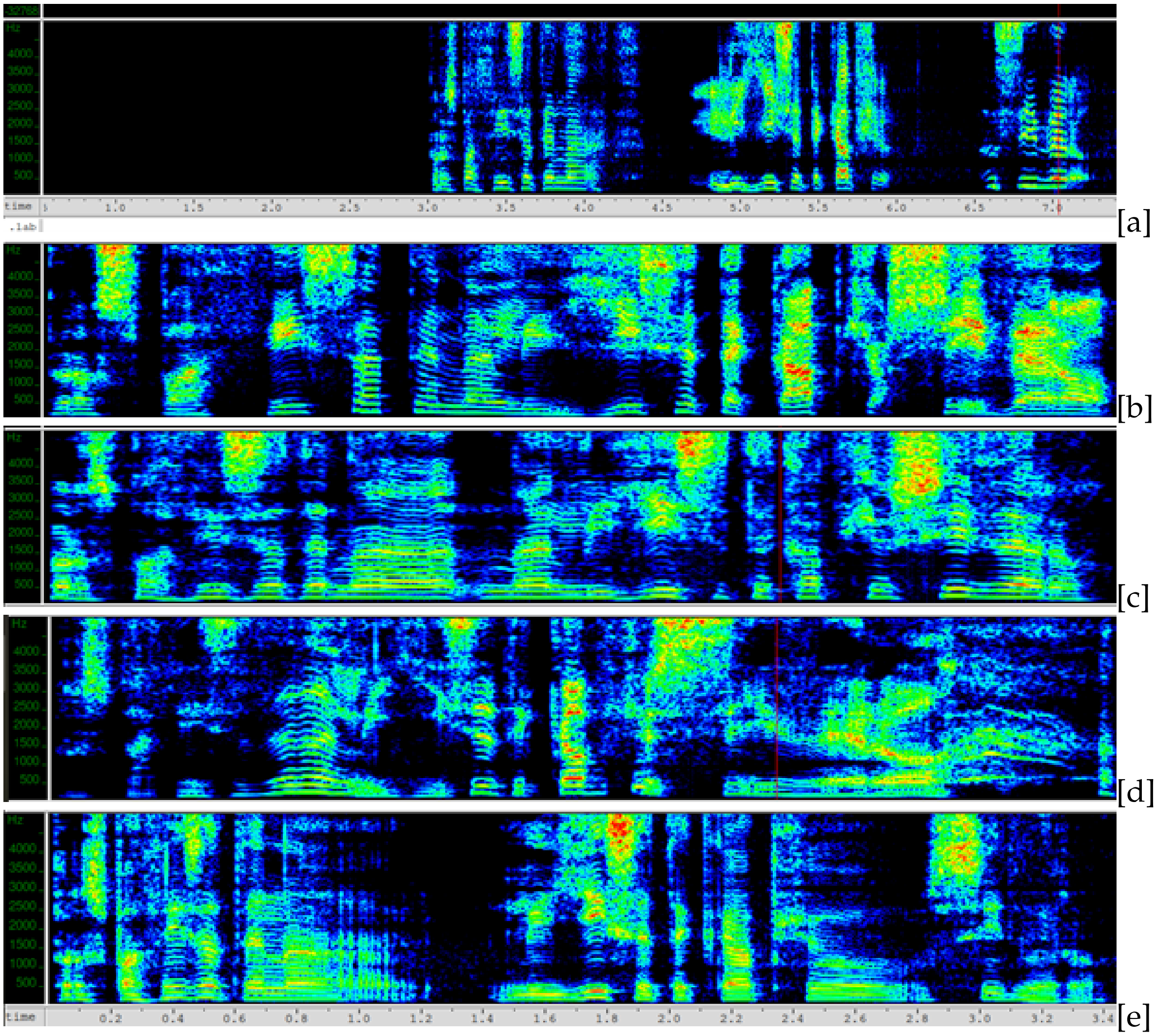


Figure 4.1: Narrowband spectrograms of 5 emotions of a male speaker: (a) disgust, (b) anger, (c) fear, (d) happiness, and (e) sadness.

counterparts as they represent pitch source harmonics clearly in the form of horizontal striations. Moreover, as discussed in sub-Section 4.1, the pitch and its harmonics (along with its dynamic implementation) are important correlates of speech prosody to convey a particular emotion in speech signal. Pauses or silences between the words, energy content, and gaps between the horizontal striations are studied to identify the difference between each emotion. Another observation from Figure 4.1 and Figure 4.2 is that the gap between pitch source harmonics is higher in Figure 4.2 than in Figure 4.1 as females have higher F_0 (fundamental frequency) than males.

4.3 Energy of Emotions

The energy measurements are obtained using the standard discrete-time energy formula. In particular, for a discrete-time signal $x(n)$, energy is computed as E=

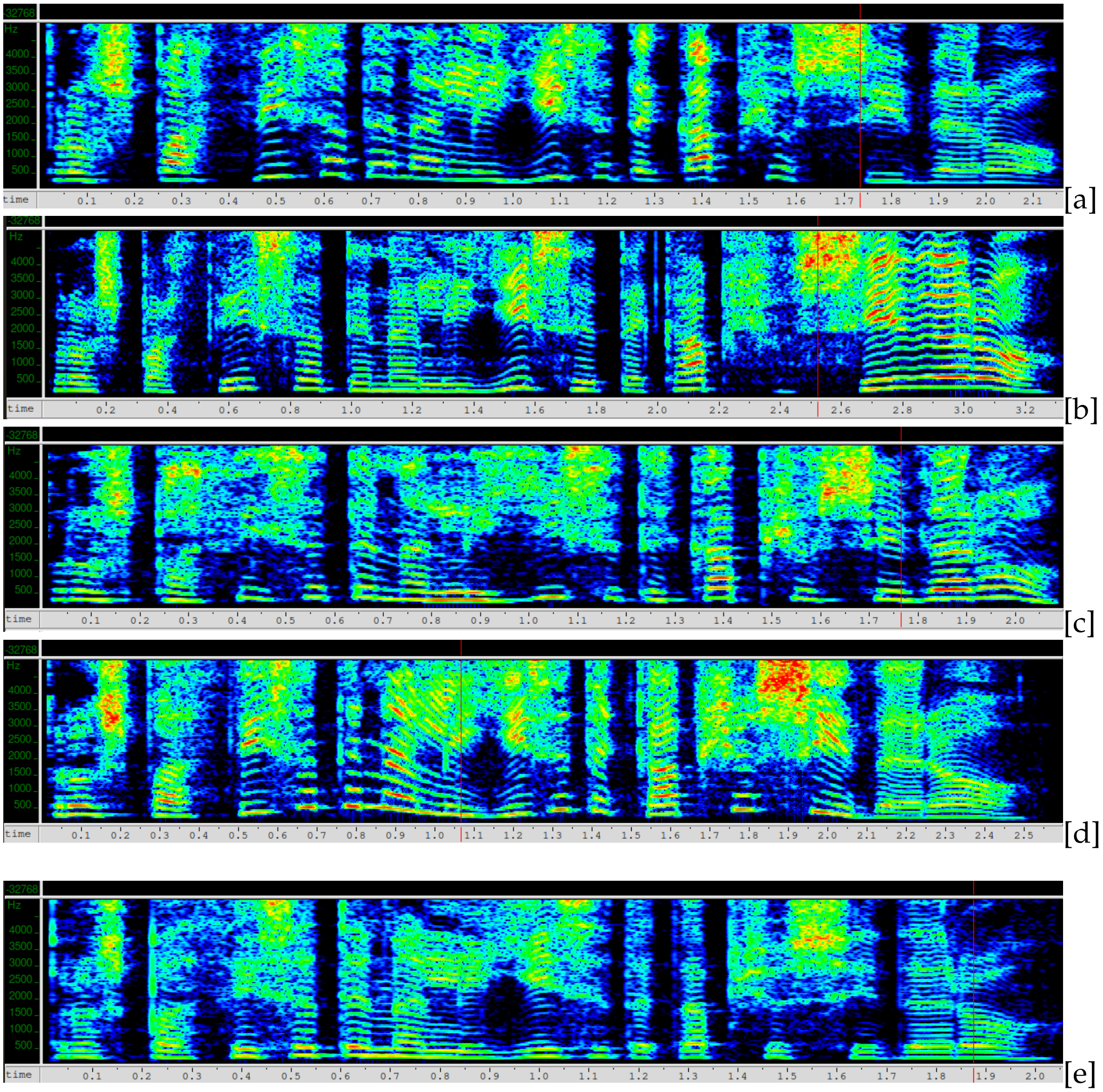


Figure 4.2: Narrowband spectrogram of 5 emotions of a female speaker: (a) anger, (b) disgust, (c) fear, (d) happiness, and (e) sadness.

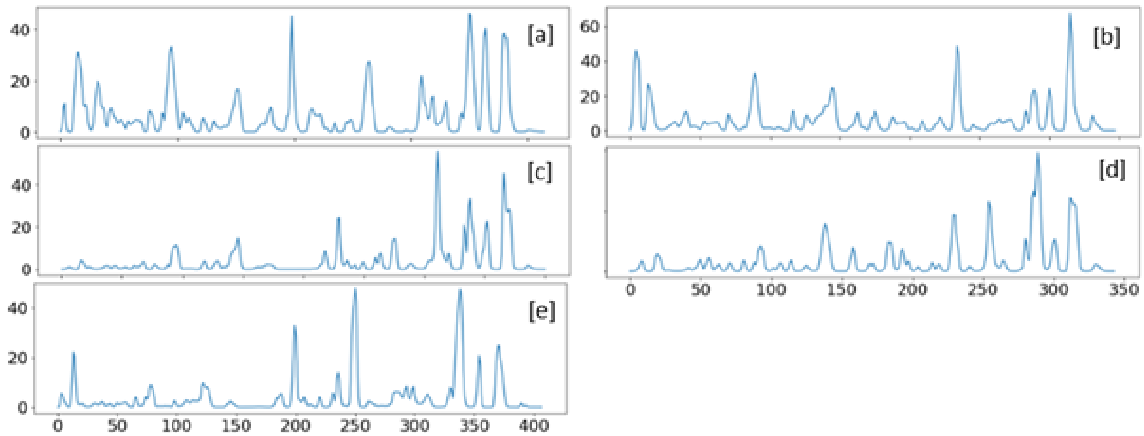


Figure 4.3: l^2 Energy of ***anger*** by 5 speakers uttering the same sentence ((a) female1, (b) female2, (c) male1, (d) male2, and (e) female3).

$\sum_{n=-\infty}^{\infty} |x(n)|^2$. It should be noted that this energy, E is conserved in frequency domain (i.e., Parseval's energy equivalence) and also in the Short Time Fourier Transform (STFT) framework, for the continuous-time version of STFT). The sampling rate of the audio files is 44.1 kHz. The audio sample is then frame-blocked and windowed (Hann window), so the acoustic events in the speech sample can be represented as a compact set of speech parameters. Hop size is taken to be 10 ms, and Fast Fourier Transform (FFT) size is 2048 samples. Observations on how energy changes for different emotions and between gender is analyzed. Figure 4.3-4.7 shows the plots of 5 speakers (three females and two males) uttering the same sentence in 5 emotions, namely, anger, disgust, happiness, fear, and sadness. Energy is then represented in the form of a boxplot. Figure 4.8[a] describes how a male speaker (1 male speaker chosen of 2 male speakers) utters a sentence in 5 emotions (disgust, anger, fear, happiness, and sadness, respectively). Figure 4.8[b] provides the same information for a female speaker (plot of 1 female speaker chosen from 3 female speakers). Boxplot provides a visual data summary, enabling us to identify the dataset's average score, skewness, dispersion, and outliers. Histogram is plotted for a male and female speaker (Figure 4.9[a] and Figure 4.9[b], respectively) to find the quantitative distribution of signal's energy.

4.4 Teager Energy Operator (TEO)

This non-linear feature or operator was introduced in 1990s by Teager and Kaiser [38]. Speech is produced by non-linear, vortex-airflow interaction in the vocal tract system. Stressful situation affects the muscle tension of the speaker that results in an alteration of the glottal airflow during the production of the sound [10]. This

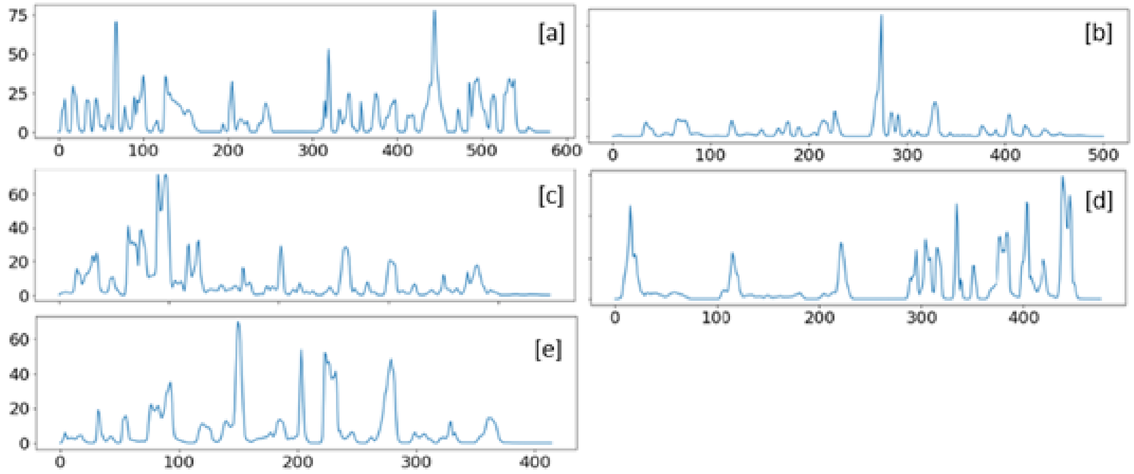


Figure 4.4: l^2 energy of ***disgust*** by 5 speakers uttering the same sentence ((a) female1, (b) female2, (c) male1, (d) male2, and (e) female3).

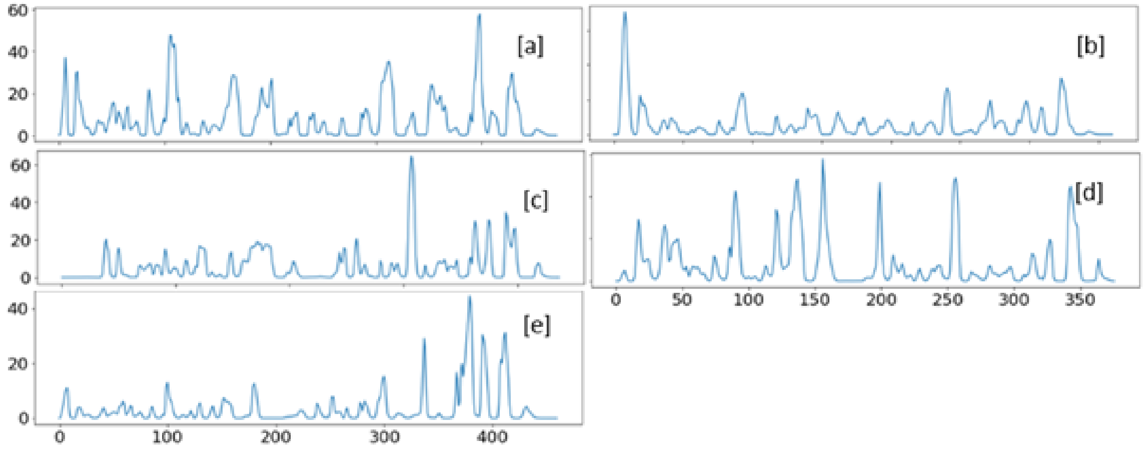


Figure 4.5: l^2 energy of ***happines*** by 5 speakers uttering the same sentence ((a) female1, (b) female2, (c) female3, (d) male1, (e) male2).

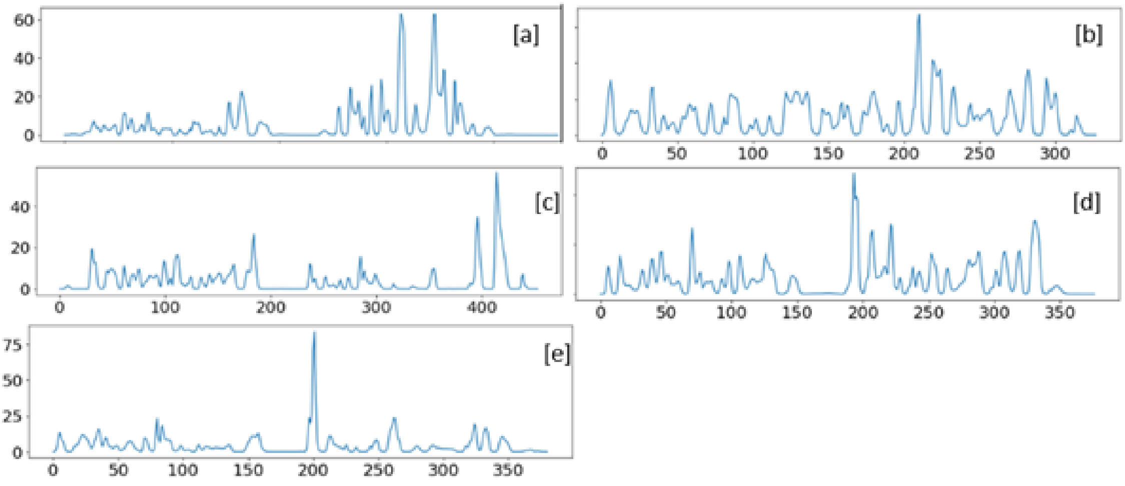


Figure 4.6: l^2 energy of ***fear*** by 5 speakers uttering the same sentence ((a) male1, (b) female2, (c) female3, (d) male2, and (e) female1).

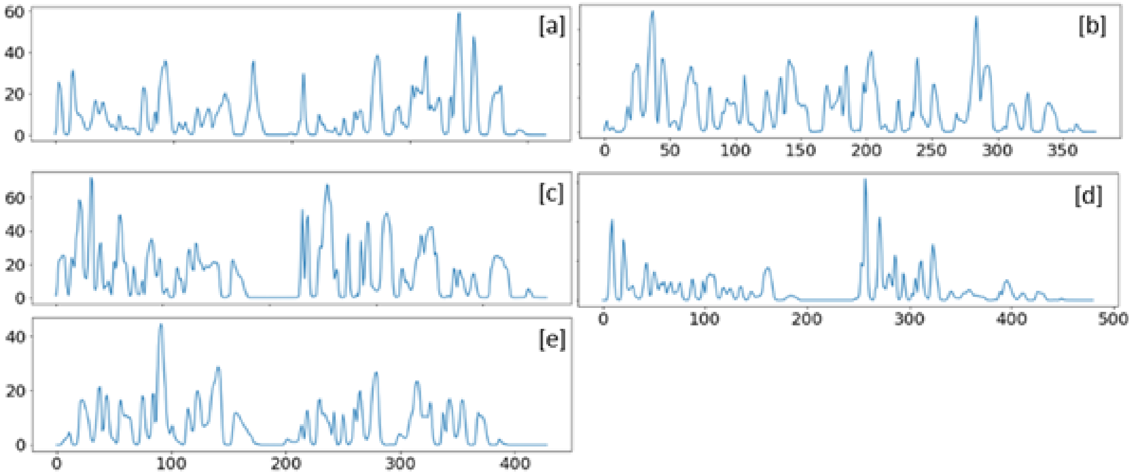


Figure 4.7: l^2 energy of ***sadness*** by 5 speakers uttering the same sentence ((a) female1, (b) female2, (c) female3, (d) male2, and (e) male1).

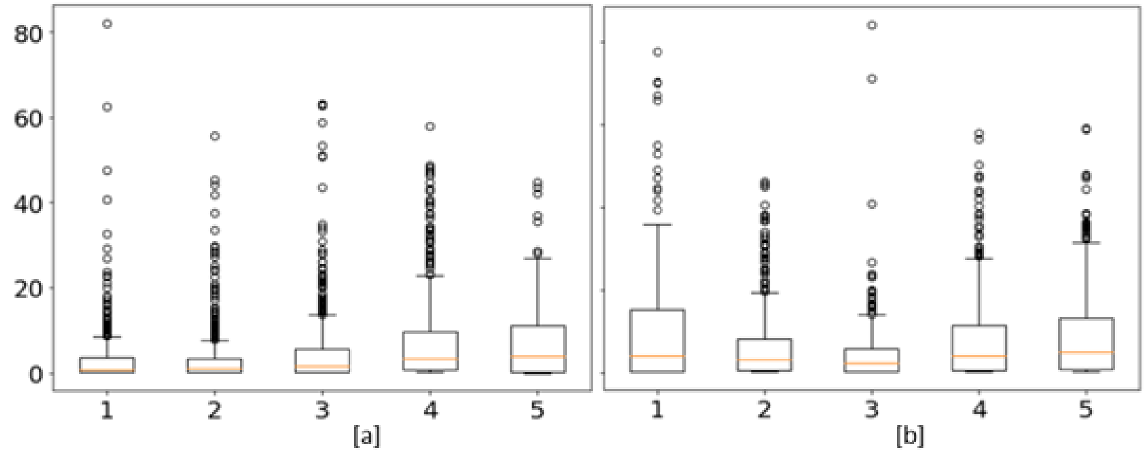


Figure 4.8: $\boldsymbol{l}^2$ energy boxplot of [a] showing male-1 speaker, and [b] showing female-1 speaking the same sentence in 5 emotions- (1) disgust, (2) anger, (3) fear, (4) happy, and (5) sadness.

is captured ***via*** TEO, in particular,

$$\Psi\{x(n)\} = x^2(n) - x(n+1)x(n-1), \tag{4.1}$$

where $\Psi\{\}$ is the Teager Energy Operator and $x(n)$ is the discrete time signal.

In Figure 4.10 and Figure 4.11, the x-axis represents the frames, and the y-axis amplitude is the TEO plots obtained for a male speaker speaking a sentence in 5 emotions and a female speaker uttering the same sentence (as a male) in 5 emotions, respectively. The energy distribution, silence, or pauses were observed to distinguish between the emotions.

4.5 Inferences

The results analyzed here take at least five sentences of each emotion; only some are plotted in this thesis. Figure 4.12 shows that the highest energy content is available in anger and the least in sadness. It is also noted that short pauses are highest in anger. This result is linked to the fact that when one gets angry, breathing is released in shorter and quicker puffs, and the force with which air is released from the lungs to the vocal tract system is also high, thus, showing these characteristics.

The amplitude of energy plots are shown in Table 1. It is found that females have higher amplitude peaks apart from the disgust emotion. This states that in females, loudness is seen more than in males while expressing emotions. We also see that fear and disgust have the highest amplitude in females and males,

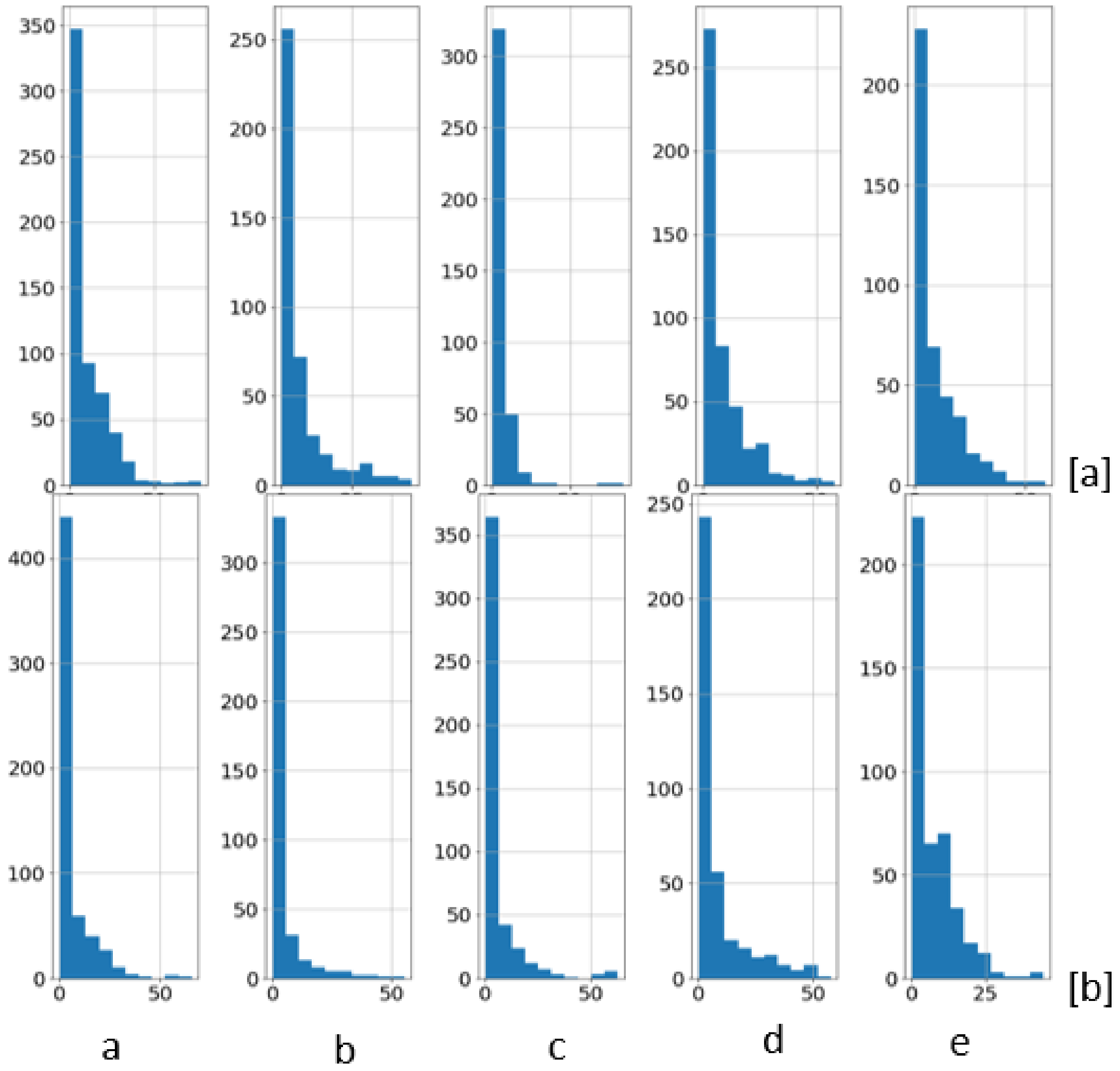


Figure 4.9: l^2 energy histogram of [a] showing male-1 speaker, and [b] showing female-1 speaking the same sentence in 5 emotions- (a) disgust, (b) anger, (c) fear, (d) happy, and (e) sadness.

respectively. Surprisingly, we observed that anger has the least amplitude of all.

Table 4.1: Highest amplitudes reached for each emotion by male and female speakers

Amplitude	**Anger**	**Disgust**	**Fear**	**Happy**	**Sad**
Females	60 dB	75 dB	80 dB	70 dB	65 dB
Males	55 dB	80 dB	60 dB	60 dB	60 dB

Boxplots of energy is shown in Figure 4.8(a) and Figure 4.8(b) for male and female, respectively. It is observed that the median is lowest for disgust and highest for sadness in males, whereas, in females, fear has the lowest median, with the other emotions having similar median values. All the emotions are positively

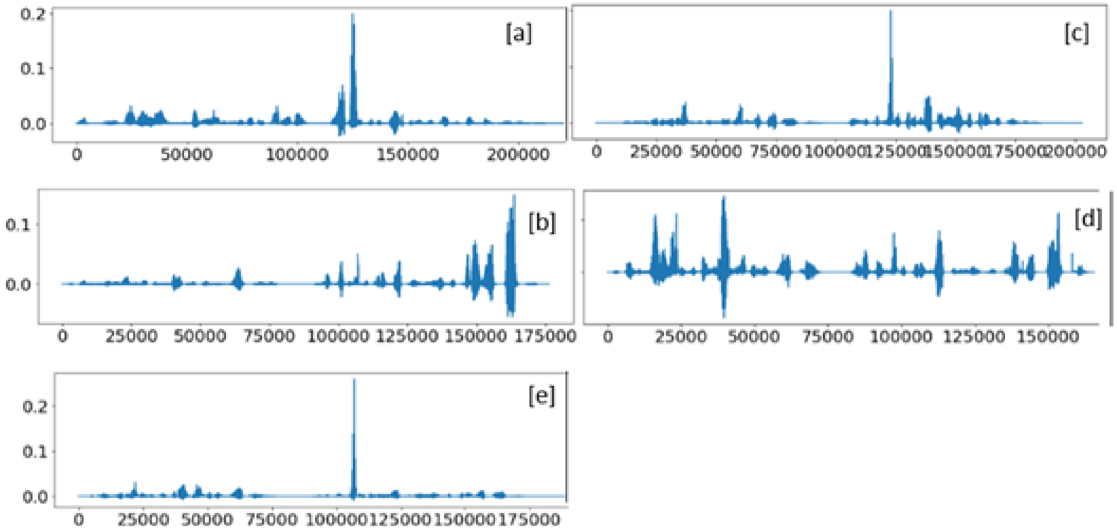


Figure 4.10: TEO profile of male 1 uttering same sentence in 5 emotions- (a) disgust, (b) anger, (c) fear, (d) happy, and (e) sad.

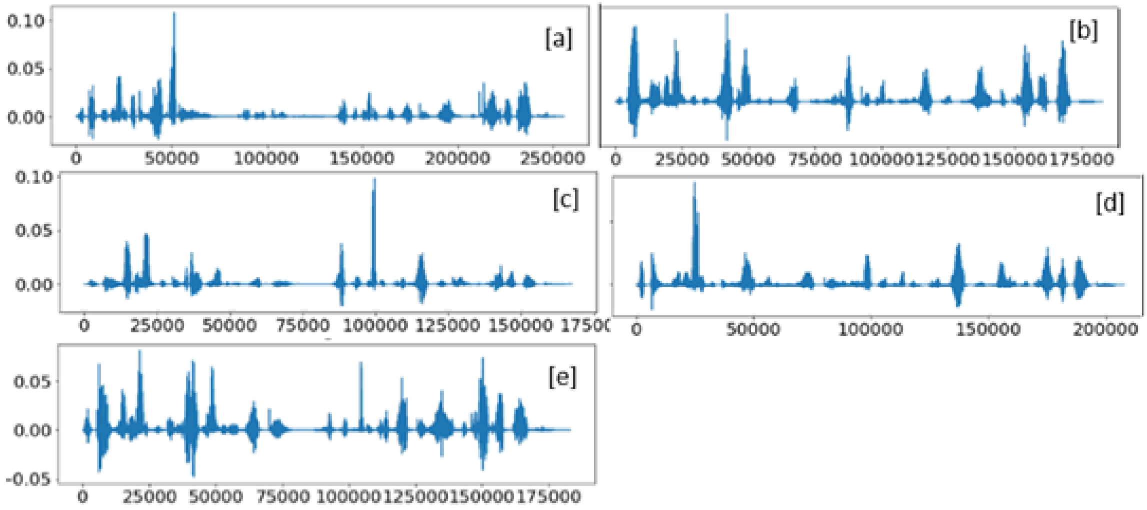
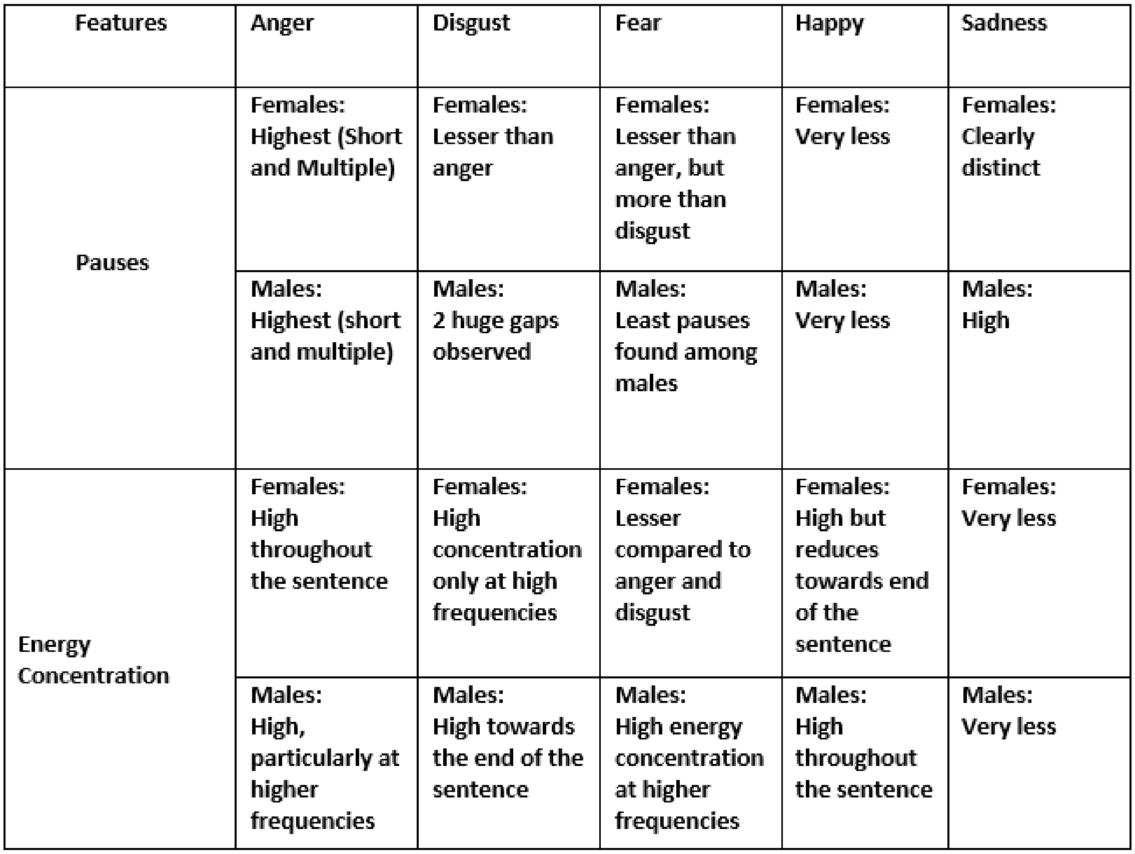


Figure 4.11: TEO profile of female 1 uttering same sentence in 5 emotions- (a) disgust, (b) anger, (c) fear, (d) happy, and (e) sad.



Features	Anger	Disgust	Fear	Happy	Sadness
Pauses	Females: Highest (Short and Multiple)	Females: Lesser than anger	Females: Lesser than anger, but more than disgust	Females: Very less	Females: Clearly distinct
	Males: Highest (short and multiple)	Males: 2 huge gaps observed	Males: Least pauses found among males	Males: Very less	Males: High
Energy Concentration	Females: High throughout the sentence	Females: High concentration only at high frequencies	Females: Lesser compared to anger and disgust	Females: High but reduces towards end of the sentence	Females: Very less
	Males: High, particularly at higher frequencies	Males: High towards the end of the sentence	Males: High energy concentration at higher frequencies	Males: High throughout the sentence	Males: Very less

Figure 4.12: Observations from Narrowband spectrograms of emotions using AESDD.

skewed, implying most values are towards the lower bound. Sadness emotion has the most even spread among all the emotions and thus, has the least outliers. Spread/ dispersion is least in fear for females and anger for males. These findings of ours are in agreement with the original study reported in [93]. This implies sadness and happiness have greater variability in speaking than the other emotions.

The histogram in Figure 4.9 is a right-skewed distribution, i.e., data values fall at the lower range (as seen in the boxplots). Most of the values occur within the interval of 0-10 dB for all the emotions, and fear has the least range of 0-20 dB, within which almost all its values are covered, proving why fear has the least median in females.

TEO plots in (Figure 4.10 and Figure 4.11) show that for male speaker 1, happiness followed by anger has the greatest energy profiles, and sadness has the lowest. However, for female speaker 1, the results obtained were surprisingly different from the other analysis, as sadness and anger have the greatest energy profiles, and fear has the lowest energy profile. This result may be because of *cognitive psychology* [31]. Studies prove that emotions are not always instantaneous rather, it is a build-up of feelings that are acquired over some time [31], [50]. This makes sense, as sometimes, a small trigger is enough to make a person very emotional, leading to extreme reactions.

4.6 Cultural and Linguistic Effects on Emotions: English *vs.* Mandarin

In this era, where the population and technology are increasing rapidly, communication among and between them is essential. Language plays its role well in human interaction as well as in human-machine interaction. Emotional Voice Conversion (EVC) is a technique to convert the emotional state of an utterance to another without changing the linguistic information and speaker's identity. Its applications are enormous in human-machine interaction, developing emotional Text-To-Speech (TTS), etc. Several languages in South-East Asia and Africa are tonal, where pitch or F_0 differences are used to differentiate meanings of words or to convey grammatical distinctions. In contrast, English is a stress language, i.e., in this language, the tone is used to convey an attitude or change a statement to a question; however, it does not affect the meaning of individual words [1]. The analysis presented in this Section is useful for conversion between languages and between emotions. Here, we analyze the loudness parameter using RMSE, voiced and unvoiced components using ZCR, and F_0 and its harmonics using narrowband spectrograms for ESD corpus.

4.6.1 Spectrographic Analysis

In this work, we study the narrowband spectrograms (as it gives good frequency resolution, i.e., shows pitch source harmonics as horizontal striations, useful for tonal language analysis), and F_0 of English and Mandarin sentences spoken in 5 emotions, namely, anger, happy, neutral, sad, and surprise. The energy distribution, pitch source harmonics, and silences are compared. Figure 4.13 and Figure 4.14 show the F_0 changes, plot, and spectrograms of female speakers uttering the same sentence in English and Mandarin, respectively. The detailed analysis of spectrograms is presented in Figure 4.15. We infer that high energy contents are seen in all five emotions of Mandarin speech and thus, indicating that Mandarin is usually louder than English. A significant difference in spectrograms is that all English sentences with five emotions had energy components present only at the higher frequency at the end of a sentence, which wasn't seen in any spectrograms for Mandarin. The width between the horizontal striations gives pitch (the way the auditory system perceives frequency) information, which is higher in Mandarin than in English. The silences were seen more in Mandarin than in English.

The study of F_0 contour is represented in the form of a boxplot (which gives the spread or variance of F_0) in Figure 4.16. It is noted that neutral emotion has the least spread in both languages, and the highest spread is seen in emotions of surprise and anger in English and Mandarin speech, respectively. Almost no outliers are seen for Mandarin speech, i.e., there is not much difference between the F_0 values as compared to the English. Another distinction seen is that the median values for all emotions in Mandarin are higher than that in English. These conclude that the F_0 contours are at higher frequencies and with wide fluctuations for Mandarin speech.

4.6.2 Root Mean Square Energy (RMSE)

RMSE for speech signal is a crucial acoustic cue for target speech perception because hearing is the process of detecting energy [90]. It is the squared signal value (amplitude), averaged over time, and its square root is calculated, as represented-

$$RMS_t = \sqrt{1/K \sum_{k=t.K}^{(t+1)(K-1)} |s(k)^2|}, \tag{4.2}$$

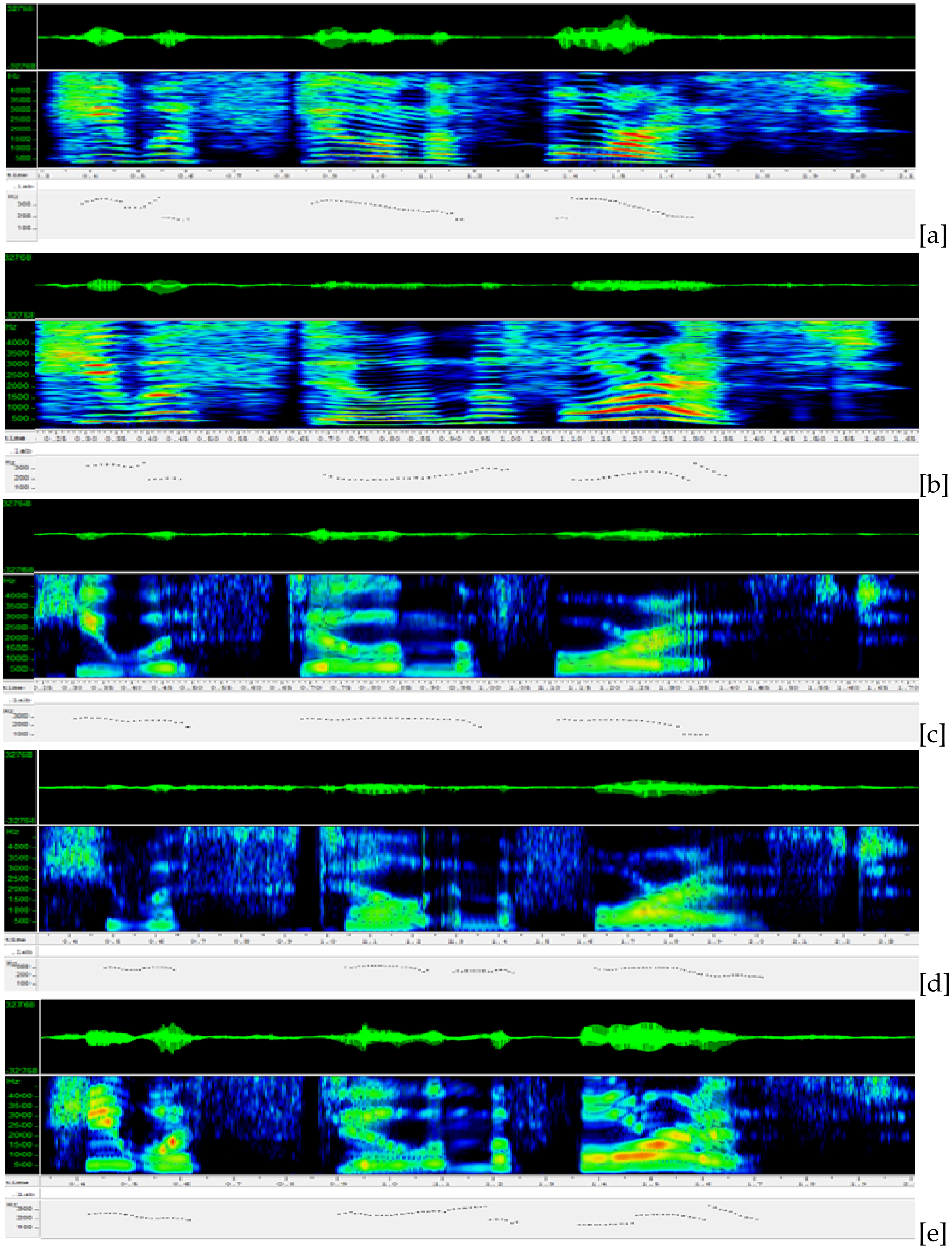


Figure 4.13: Time-domain signal, narrowband spectrograms, F_0 contour of **English** sentences by female speakers from ESD corpus for five emotions: (a) anger, (b) happy, (c) neutral, (d) sad, and (e) surprise.

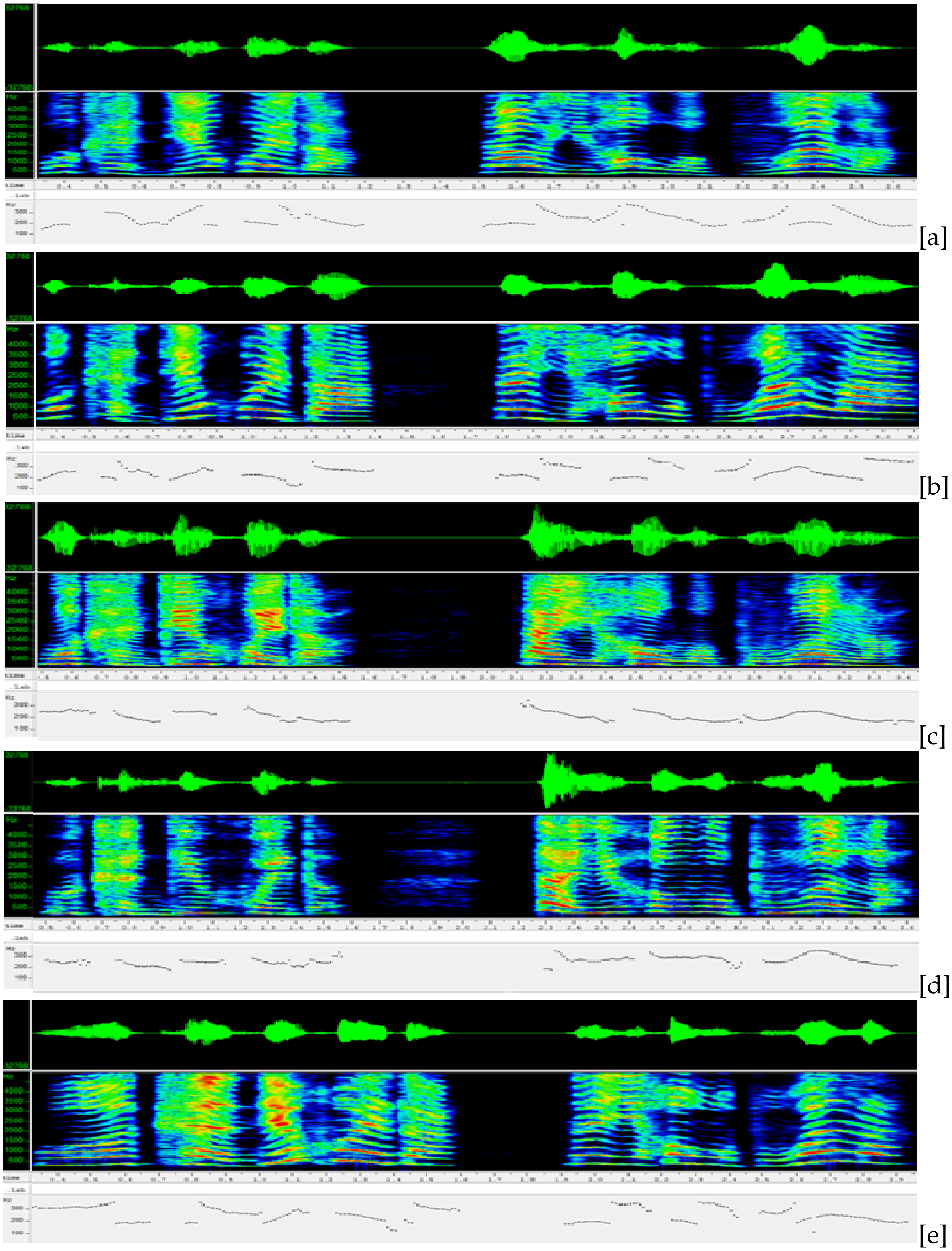
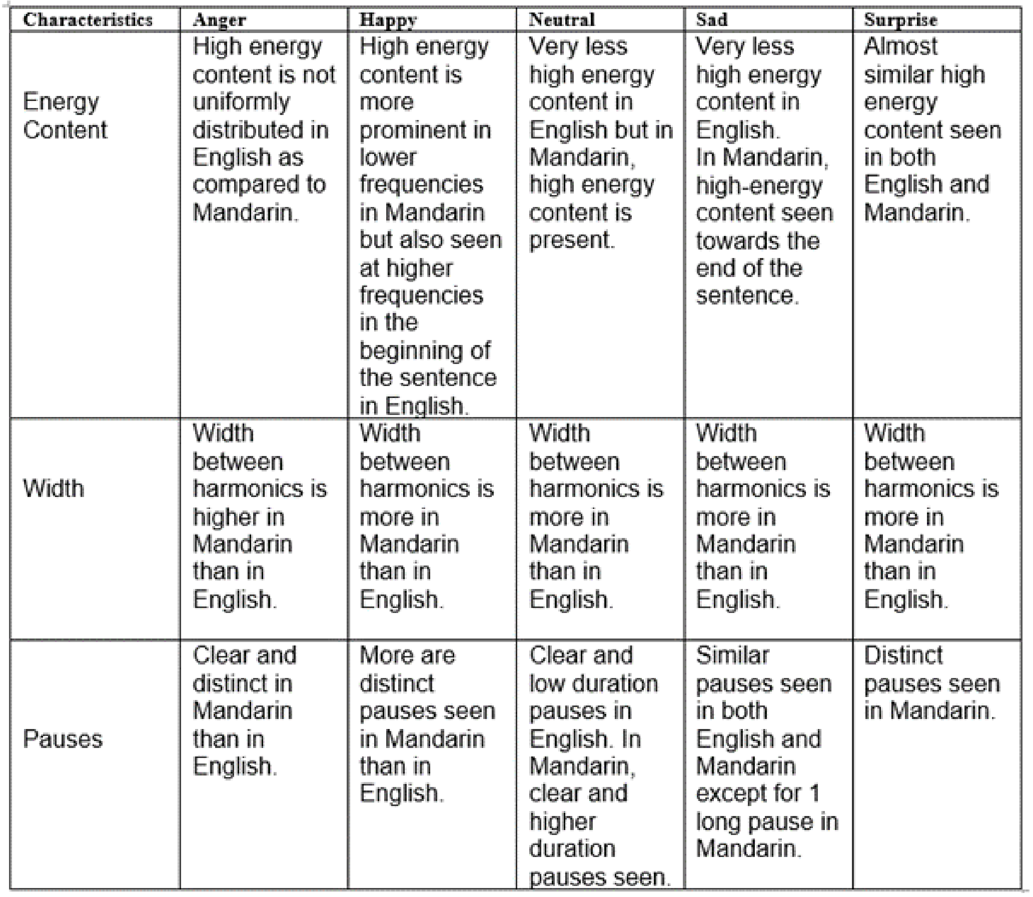


Figure 4.14: Time-domain signal, narrowband spectrograms, F_0 contour of **Mandarin** sentences by female speakers from ESD corpus for five emotions: (a) anger, (b) happy, (c) neutral, (d) sad, and (e) surprise.



Characteristics	Anger	Happy	Neutral	Sad	Surprise
Energy Content	High energy content is not uniformly distributed in English as compared to Mandarin.	High energy content is more prominent in lower frequencies in Mandarin but also seen at higher frequencies in the beginning of the sentence in English.	Very less high energy content in English but in Mandarin, high energy content is present.	Very less high energy content in English. In Mandarin, high-energy content seen towards the end of the sentence.	Almost similar high energy content seen in both English and Mandarin.
Width	Width between harmonics is higher in Mandarin than in English.	Width between harmonics is more in Mandarin than in English.	Width between harmonics is more in Mandarin than in English.	Width between harmonics is more in Mandarin than in English.	Width between harmonics is more in Mandarin than in English.
Pauses	Clear and distinct in Mandarin than in English.	More are distinct pauses seen in Mandarin than in English.	Clear and low duration pauses in English. In Mandarin, clear and higher duration pauses seen.	Similar pauses seen in both English and Mandarin except for 1 long pause in Mandarin.	Distinct pauses seen in Mandarin.

Figure 4.15: Analysis of narrowband spectrograms for English ***vs.*** Mandarin emotions taken from ESD corpus.

where $s(k)^2$ is the energy of k^{th} sample, then we sum the energies of all the samples at time $\boldsymbol{t}$. To get the mean, it is then divided by frame size, K.

This feature has significant applications in audio segmentation and music genre classification. In this work, we plot the RMSE values of audio to find the loudness measure. Amplitude Envelope (AE) can also measure loudness; however, RMS is preferred as it is less sensitive to outliers than AE. In addition, it gives us perceived loudness, i.e., how our ear perceives loudness. In Figure 4.17, each plot depicts the RMSE values of the same sentences spoken in English (yellow colored) and Mandarin (Red colored) by two female (1 for English and 1 for Mandarin) speakers. From this, we observe that all the emotional sentences spoken in Mandarin have significant fluctuations in peaks compared to the English statements. Anger and surprise emotions have similar peaks in both the languages. Neutral and sad sentences in English have almost no variations in peaks. Happy in Mandarin has broader peaks. These results state that Mandarin sentences are perceived as louder (as they have more energy content, as seen from spectrograms)

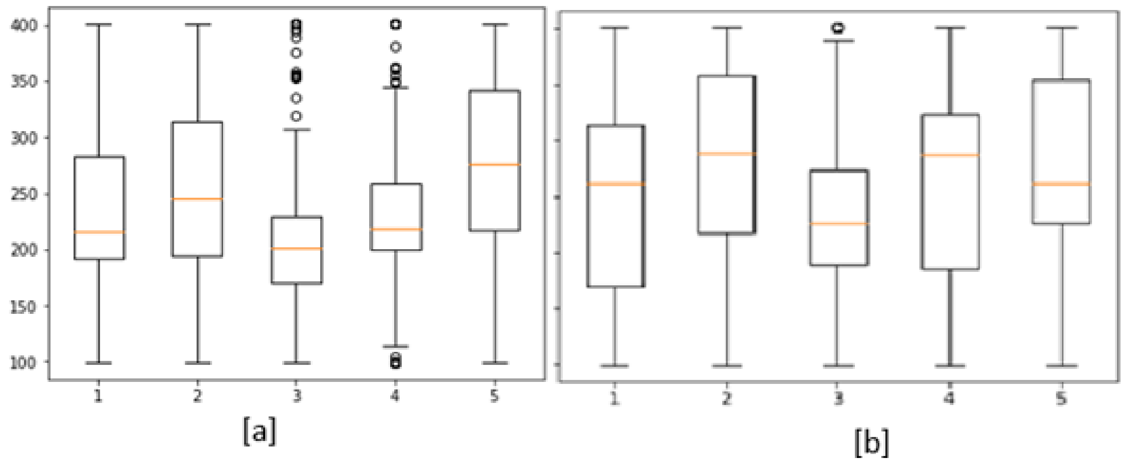


Figure 4.16: Boxplot of F_0 contour of female speaker uttering an [a] English and [b] Mandarin sentence taken from ESD corpus for [1] anger, [2] happy, [3] neutral, [4] sad, and [5] surprise.

than the corresponding English sentences.

4.6.3 Zero-Crossing Rate

ZCR is "the rate at which a signal changes from positive-to-zero-to-negative or from negative to zero to positive." Historically, it is known to correlate with formants and thus, helpful for speech perception [48]. It is expressed as-

$$ZCR_t = (1/2) \sum_{k=t.K}^{(t+1)(K-1)} |sgn(s(k)) - sgn(s(k+1)), \tag{4.3}$$

where s(k) and s(k+1) represent the amplitude at sample k and its consecutive amplitude sample, respectively. It is an useful measure to recognize percussive (random ZCR) ***vs.*** pitched sounds (stable ZCR) [18]. For this work, we use ZCR for monotonic pitch estimation and for analysing the voiced and unvoiced segments of an audio signal [14]. Figure 4.18 shows the ZCR plot for two females (1 for English and 1 for Mandarin) speaking the same sentence in both languages with five emotions: anger, happy, neutral, sad, and surprise, respectively. We can consider two extreme cases of spectral energy density, i.e., the low frequency and high frequency regions. It is observed that ZCR peaks are less in lower and high in higher frequency regions of spectrograms. ZCR peaks of Mandarin are less than that of English as tonal sounds are pitch-dependent and have voiced speech as compared to English, which has unvoiced and whisper elements (at the beginning of the sentence, as shown in Figure 4.18 for the sentence analyzed, and thus, proving that ZCR peaks are high for unvoiced sounds in comparison to voiced

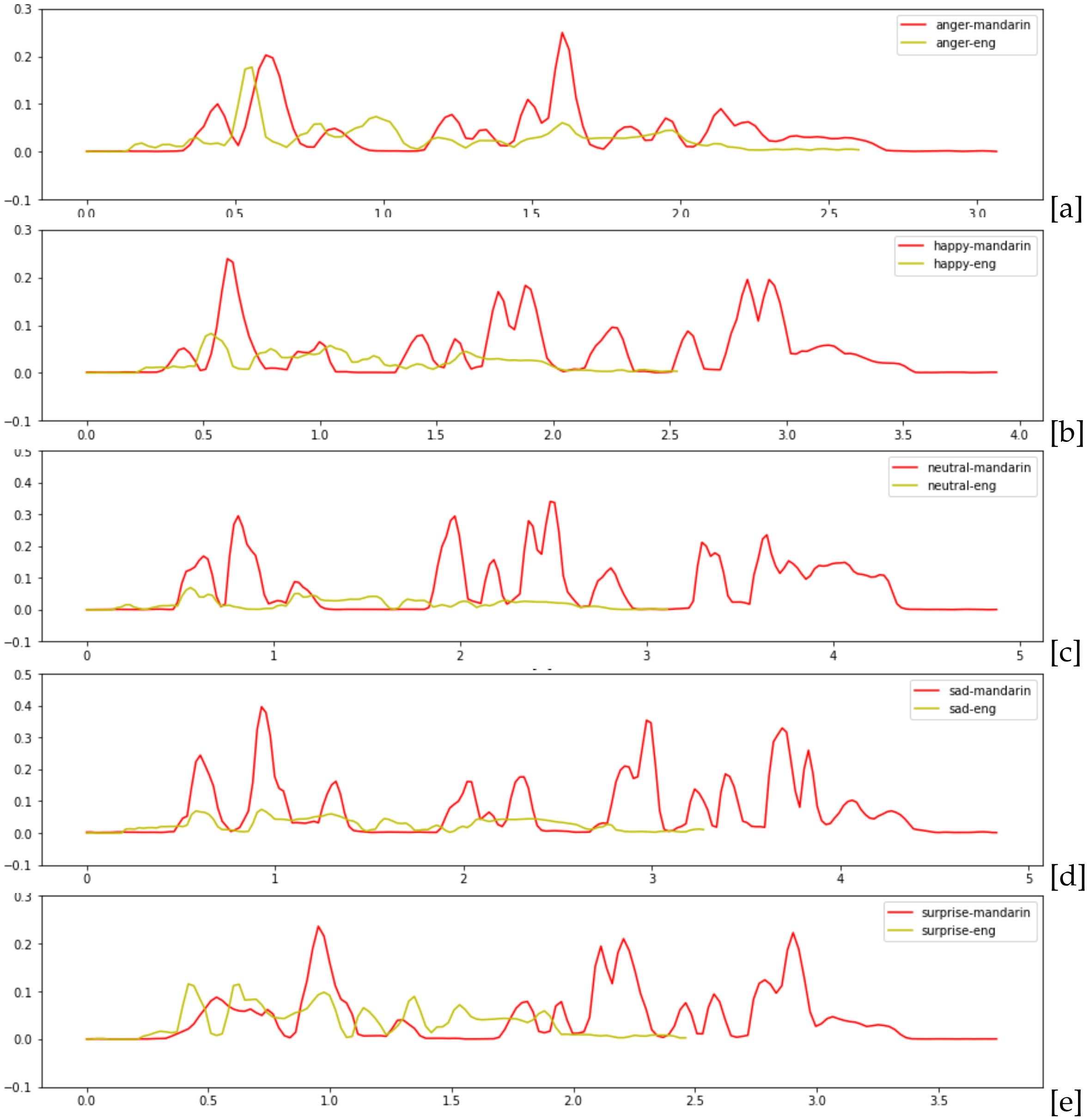


Figure 4.17: RMS for Mandarin *vs.* English for a sentence in [a] anger, [b] happy, [c] neutral, [d] sad, and [e] surprise by female speakers of ESD corpus.

sounds).

4.6.4 TEO Profile

Figure 4.19 and Figure 4.20 have the TEO profile of a female speaker uttering the same sentence with five emotions in English and Mandarin, respectively, with the X-axis representing frames and the Y-axis amplitude. These plots show that Mandarin sentences have higher energy profiles (peaks reach higher amplitudes) than English sentences. This is because a higher pitch leads to higher loudness and, thus, higher amplitude.

4.7 Chapter Summary

This study investigated five emotions: anger, fear, disgust, happiness, and sadness. This is carried out by using prosodic features of speech, in particular, loudness (amplitude measure), energy and pauses or silences, and narrowband spectrograms. We also use the novel TEO on AESDD dataset to get energy profiles. It is found that anger and happiness show similar spectrographic characteristics, however, it can be distinguished by plotting their energy spectrum. Sadness has the most spread and balanced data values among the other emotions. Characteristics of disgust seemed very similar to anger. Distinct differences between male and female emotions can also be seen (boxplot- Fig.8) by acquiring their energy values. Along with speech features, cognition's impact is equally important for recognizing a particular emotion. We then analyzed a tonal language (Mandarin) and a stress language (English) using prosodic features, such as energy, F_0, loudness, and TEO-based features. Our analysis indicates the Mandarin language has higher F_0 fluctuations due to variations in pitch, is louder, and has higher energy profiles than English. Therefore, for EVC, RMS and ZCR features can be used to maintain the speaker's identity. Analyzing how RMS and ZCR features would work if replaced with F_0 in the baseline paper [94] for EVC would be interesting. In the next chapter, we propose a new feature based on prosodic features for SER.

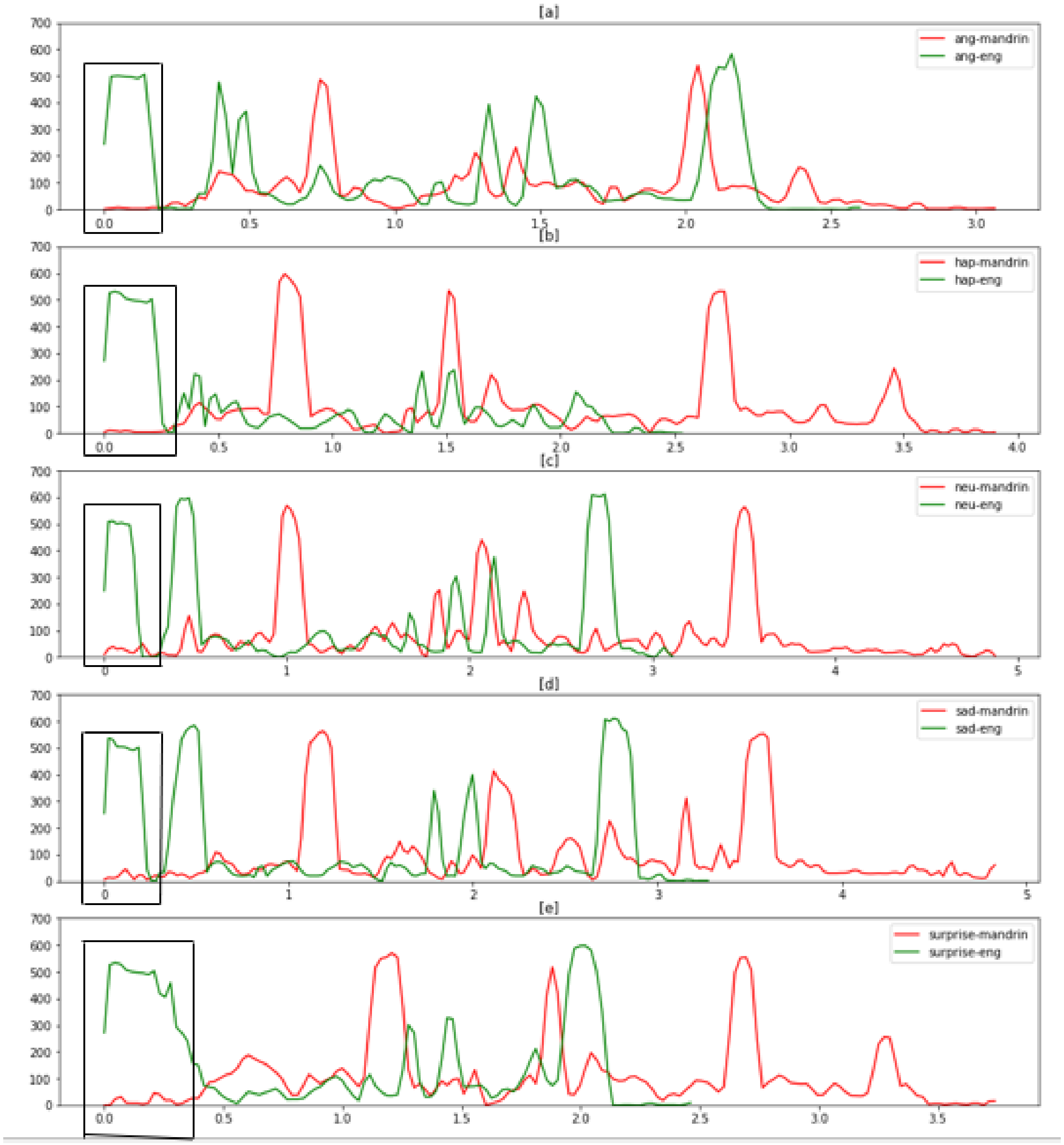


Figure 4.18: ZCR for Mandarin *vs.* English for a sentence from ESD corpus in [a] anger, [b] happy, [c] neutral, [d] sad, and [e] surprise by female speakers. The box at the beginning of the plot indicates the whisper sound |h| in "he" uttered.

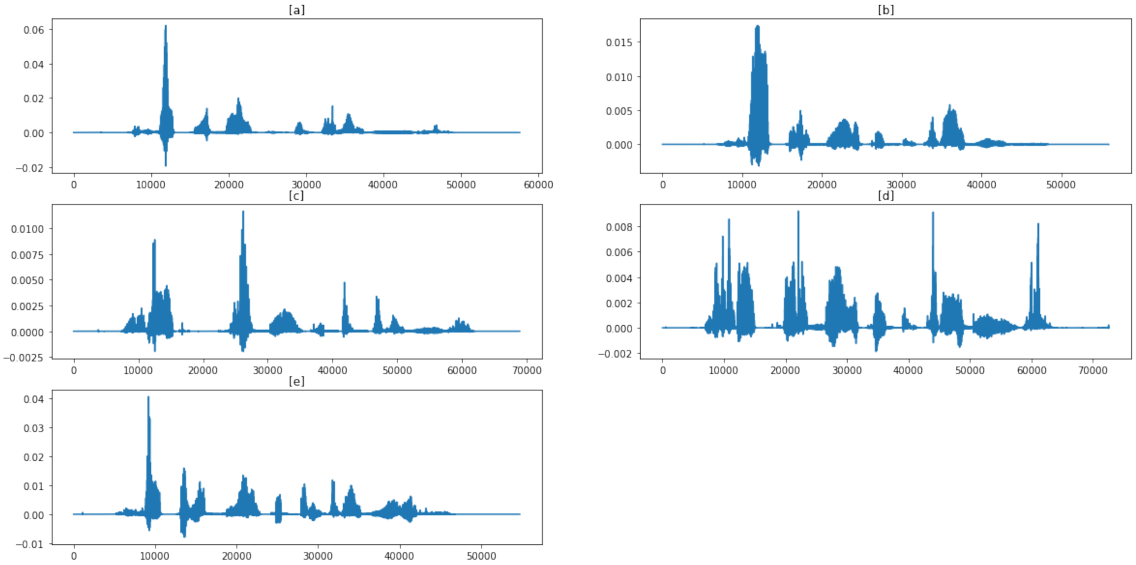


Figure 4.19: TEO profile of a female speaker uttering an English sentence from ESD corpus in [a] anger, [b] happy, [c] neutral, [d] sad, and [e] surprise.

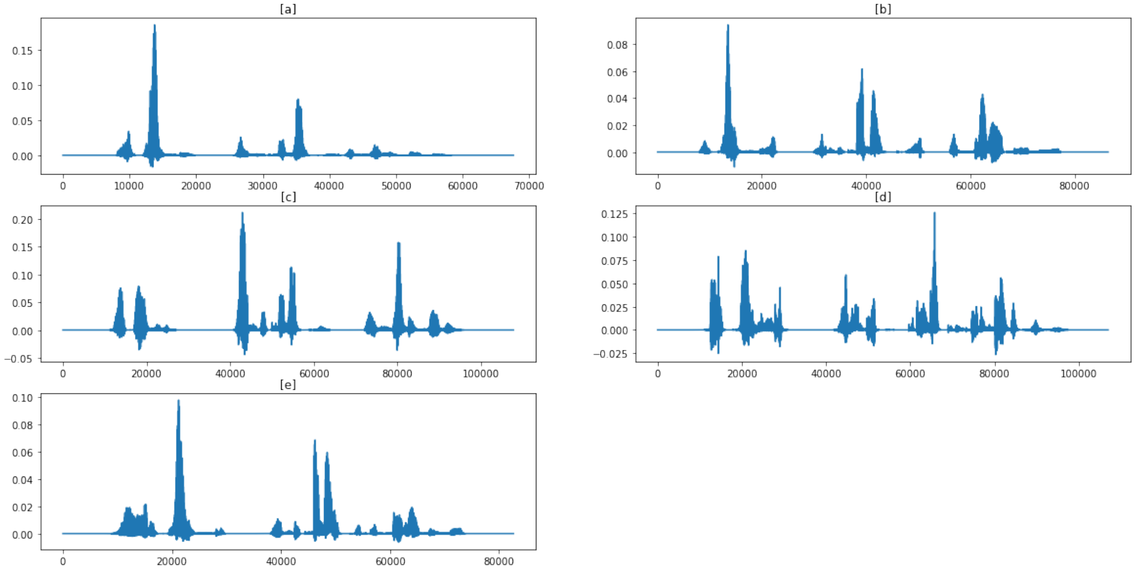


Figure 4.20: TEO profile of a female speaker uttering a Mandarin sentence from ESD corpus in [a] anger, [b] happy, [c] neutral, [d] sad, and [e] surprise.

CHAPTER 5

Music Motivated Features for SER

Prosodic features are the most commonly used feature for SER task. This chapter proposes prosodic features (high resolution pitch information and high-resolution harmonic information) motivated from music research for SER.

5.1 Introduction

Speech is the most natural and powerful form of communication. As emotions play a significant role in communication, detecting and analyzing the same is vital. However, to develop an efficient SER, it is very essential to understand the acoustics of various emotions. In particular, humans have a remarkable ability to convey and perceive emotions in speech using various prosodic cues, such as loudness (amplitude), pitch, duration of speech sound units, intonation (i.e., derivative of pitch contour), and energy of speech wave. The existing SER literature focuses on lower frequencies of speech signals for emotion recognition [78], which is why a non-linear time-frequency representation is needed. The Constant Q Transform (CQT) offers higher frequency resolution in low-frequency regions and higher time resolution in high frequency regions. It has been found that music arouses emotion [71]. The timbral textures, rhythmic contents, melody, and pitch components are parameters for detecting emotions in music [47]. Music emotion recognition is the process of identifying the emotions evoked by lyrics and melodies in music [41]. These original findings motivated us to exploit music-based features for SER.

In particular, we propose the use of CQT-based features, namely, Constant Q Harmonic Coefficients (CQHC), Constant Q Pitch Coefficients (CQPC), and combined features (CQHC + CQPC), which are shown to be efficient timbre-based features [66], in emotion recognition using CNN classifier with *10*-fold cross-validation for the classification of 5 emotions, namely, anger, neutrality, happiness, sadness, and surprise. Commonly used MFCC and GFCC features along with already explored CQT and Constant Q Cepstral Coefficients (CQCC) are used for performance comparison with proposed features. The ESD and TESS datasets are used for this work. The octave resolution used for CQT-based features is 14, the number of coefficients is 20, and the minimum frequency is 30 Hz.

5.2 Constant Q Transform (CQT)

Due to CQT's capacity to simulate an equal-tempered frequency scale, it was initially used in western music for music analysis [19]. Discrete Fourier Transform (DFT) uses the same window length for every frequency bin and thus, giving linear frequency resolution. CQT modifies the Short-Time Fourier Transform (STFT) such that frequencies are logarithmically spaced. Its window length decreases with increasing frequency. The ratio of the center frequency to bandwidth is *quality* factor (Q), which is kept constant and thus, the name " Constant Q". This gives log frequency resolution, where frequency bins correspond to tones. Let *x(n)* be the discrete-time speech signal obtained with a sampling frequency of *Fs*. The STFT of *x(n)* is given by [19]:

$$X(\omega,\tau) = \sum_{n=-\infty}^{\infty} x(n)\mathsf{h}(n,\tau)e^{-j\omega n}, \tag{5.1}$$

where $\mathsf{h}(n,\tau)$ represents the analysis window, centered at time τ. The window is a function of time parameter τ alone. Now, let $z(n)$ represent a frame of the speech signal, then the DFT, $Z(k)$, of the $z(n)$ can be represented as:

$$Z(k) = \sum_{n=0}^{N-1} z(n)e^{-j(\frac{2\pi}{N})kn}, \tag{5.2}$$

where k is the frequency bin index, and $\omega_{DFT} = (2\pi k)/N$. The CQT of a signal $z(n)$ is given by [19]:

$$Z^{CQT}(k) = \frac{1}{N(k)} \sum_{k=0}^{N(k)-1} z(n)\mathsf{h}(n,k)e^{-j\left(\frac{2\pi}{N(k)}Qn\right)}, \tag{5.3}$$

where $\omega_{CQT} = (2\pi Qn)/N(k)$, and $\mathsf{h}(n,k)$ is analysis window, which remains constant for each frequency bins f_k, however, its length is determined by $N(k)$ and thus, it is a function of both n (time) and k (frequency), where $N(k) = Q(F_s/f_k)$. The *quality factor* (Q) is a ratio of the center frequency to bandwidth, and it is given by [19]:

$$\because Q = \frac{f_k}{\Delta f_k} = \frac{f_k}{f_{k+1} - f_k} = \frac{1}{2^{1/B} - 1}, \tag{5.4}$$

where B represents the number of bins per octave, and f_k shows the frequency of k^{th} spectral component, which is given as:

$$f_k = (2^{(k-1)/B}) f_{min}, \tag{5.5}$$

where f_{min} is the minimum frequency of the signal. CQT can be decomposed into energy-normalized pitch components (CQPC), and pitch-normalized spectral components (CQHC).

5.3 Proposed Work

5.3.1 Constant Q Harmonic Coefficients (CQHC)

The CQT's logarithmic resolution enables the harmonics to form a stable pattern in the frequency-domain while maintaining their relative position w.r.t. F_0 [66],[19]. Constant Q Harmonic Coefficients (CQHC) is associated with timbre (quality of sound produced by voice or instrument) as harmonics carry spectral information of speech signal. As timbre is ideally independent of pitch and thus, normalizing pitch gives efficient timbre feature set. Pitch normalization is achieved by assuming that the CQT spectrum can be represented as a convolution between a pitch-normalized spectral component, and energy-normalized pitch component [66]. In particular,

$$X = S * P, \tag{5.6}$$

where X represents the CQT spectrum, S represents the pitch-normalized spectral component, and P represents the energy-normalized pitch component. From the property that the magnitude of the Fourier transform is *shift-invariant*, the spectral component can be approximated by the magnitude Fourier transform of the CQT spectrum. The IFFT of this approximation gives the estimate of the spectral component as stated in eq. (5.7) [66]:

$$S = \mathcal{F}^{-1}(|\mathcal{F}(X)|), \tag{5.7}$$

where $\mathcal{F}^{-1}(\cdot)$ represents the inverse Fourier transform. Given the octave resolution considered for the computation of CQT, we can obtain the locations of harmonics in the spectral component, and then extract the harmonic coefficients. The coefficients from the spectral component are obtained by [66]:

$$i = round(O_r log_2(k)), \tag{5.8}$$

$$CQHC_k = S(i), \tag{5.9}$$

where k takes the value between 1 and N_c, O_r is the octave resolution, and N_c is the number of desired coefficients. The CQHC captures the harmonics information of the speech signal embedded in the CQT spectrum.

5.3.2 Constant Q Pitch Coefficients (CQPC)

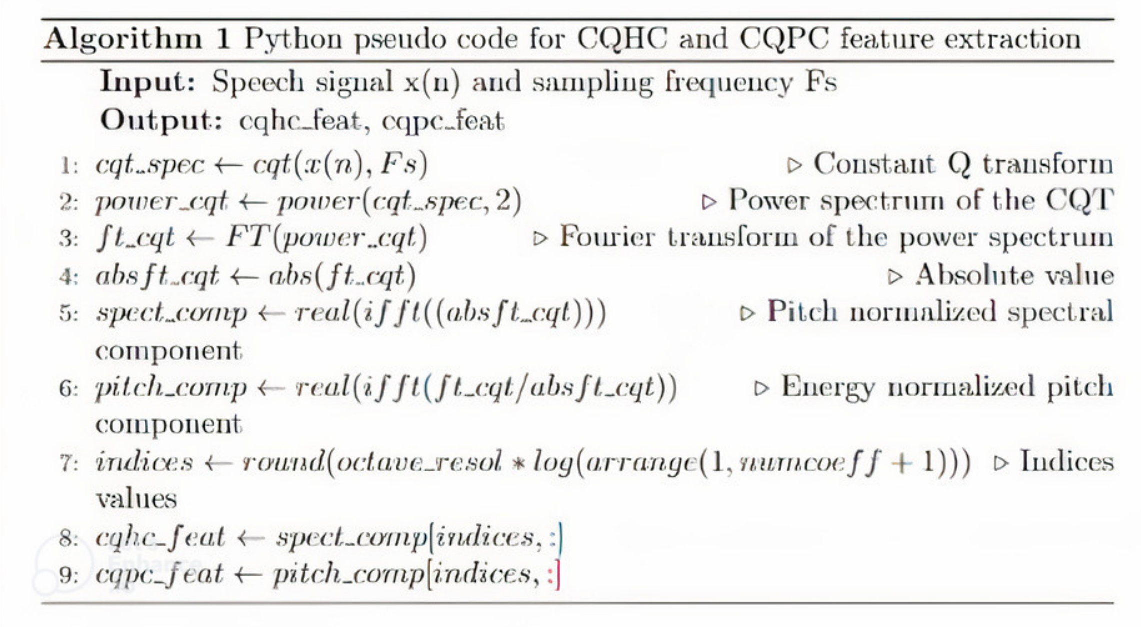
The decomposition of the CQT spectrum also results in an energy-normalized pitch component. This means that the information embedded in the fundamental frequency (F_0), and first few formants is stored through the pitch component. The pitch component is calculated as [67] :

$$C = \mathcal{F}^{-1}(e^{jArg(\mathcal{F}(A)}). \tag{5.10}$$

Constant Q Pitch Coefficients (CQPC) features are stripped down version of its energy components, leaving only the fundamental frequency of notes. The FT of pitch component (P), i.e., F{P} is the phase component of CQT spectrum, and IFT of which will give the desired pitch component. Taking the binary log and rounding the value (eq. (5.8) and (5.9)) gives the CQPC. The algorithm for CQT-based features is shown in Algorithm1.

5.3.3 Feature-Level Fusion of CQHC and CQPC

This feature set is formed by combining CQHC and CQPC. The goal is to improve emotion classification by combining high resolution timbral information (from CQHC) with pitch information (from CQPC). Both their matrices are concatenated to make a 3-dimensional matrix, or a 2D image with two channels, and sent as an input to the CNN classifier. Since both features (i.e., CQPC and CQHC) are a part of CQT, they are of the same size and hence, no padding or truncation of data is required, making the feature easier to fit in the model. This repository, [4], contains all the codes needed to extract the aforementioned features from audio.



Algorithm 1 Python pseudo code for CQHC and CQPC feature extraction

```
Input: Speech signal x(n) and sampling frequency Fs
Output: cqhc_feat, cqpc_feat
$cqt\_spec \leftarrow cqt(x(n), Fs)$    ▷ Constant Q transform
$power\_cqt \leftarrow power(cqt\_spec, 2)$    ▷ Power spectrum of the CQT
$ft\_cqt \leftarrow FT(power\_cqt)$    ▷ Fourier transform of the power spectrum
$absft\_cqt \leftarrow abs(ft\_cqt)$    ▷ Absolute value
$spect\_comp \leftarrow real(ifft((absft\_cqt)))$    ▷ Pitch normalized spectral component
$pitch\_comp \leftarrow real(ifft(ft\_cqt/absft\_cqt))$    ▷ Energy normalized pitch component
$indices \leftarrow round(octave\_resol * log(arrange(1, numcoeff + 1)))$    ▷ Indices values
$cqhc\_feat \leftarrow spect\_comp[indices, :]$
$cqpc\_feat \leftarrow pitch\_comp[indices, :]$
```

5.4 Experimental Details

ESD and TESS dataset were used for analysis from which we used *five* emotions, namely, anger, happiness, neutral, sadness, and surprise (common between datasets). The analysis is limited to English as cultural and linguistic differences impact emotions [27]. Two types of comparison among audio data are performed. One includes the parallel comparison, where both datasets are individually split in 80-20 ratio for training and testing, respectively, and a non-parallel comparison using ESD is made where Leave One Speaker Out (LOSO) method is used to remove one speaker out for testing and keeping others for training to get speaker-independent results. For non-parallel comparison, one male speaker was left out for test and others were given for training. Classifier details are given in Chapter 3. In this work, the performance of proposed feature sets is compared with two state-of-the-art features, namely, MFCC and GFCC. *13*-D coefficients were extracted keeping window length and hop length to the default parameter setting in Keras. CQCC, derived from taking log and DCT of CQT is used to compare the performance of proposed features on speaker-independent data as CQCC is proven to work well in anti-spoofing literature for the same [83]. The octave resolution is taken at 14, the number of coefficients is 20, and the minimum frequency considered is 30 Hz.

5.5 Experimental Results

5.5.1 Results Obtained from Baseline Features

Table 5.1 reports the test accuracy obtained using different features. For parallel data analysis, GFCC outperforms other baseline features, i.e., MFCC, CQCC, and CQT by ***2.11 %***, ***2.91 %***, and ***2.4 %***, respectively, in ESD dataset. The high performance can be expected because they are better aligned to capture the motion of basilar membrane in cochlea during hearing process and thus, they can better model the physical changes within the ear during hearing [39], [49]. It is observed that they outperform the proposed features on TESS dataset, but the proposed features (CQPC and CQHC+CQPC) give consistently high accuracy across datasets (refer Table 5.1). For non-parallel data, CQCC is the best performing baseline features as expected [83].

5.5.2 Performance on Parallel Dataset

The proposed CQPC gives the highest accuracy for ESD dataset (***99.51%***) and comparable accuracy in TESS (***99.60%***). The pitch component obtained from CQT gives energy-normalized pitch (resolved F_0), melody, and rhythm information, each being an implicit function of prosody (the most used features for SER). The pitch harmonics and lower formants residing in lower regions of the speech spectrum are well captured by CQPC compared to the baseline. English being a stress-timed language is dependent on rhythm [12], making the proposed features (i.e., CQPC and CQPC+CQHC) work better.

The proposed CQHC feature gives the least accuracy for ESD corpora (refer to Table 5.1). CQHC captures timbre information, which is independent of pitch and loudness (characteristics important for emotion recognition). Unlike musical instruments, the human voice is not a pure tone, but a mixture of fundamental and higher frequencies (upper harmonics), and getting resolved harmonic coefficients make CQHC give good comparable results, but since the pitch is normalized, i.e., fundamental frequency (F_0) is brought to lowest frequency bin, it gives the least accuracy as all other features capture pitch information to some extent and thus, showing the importance of pitch in emotion recognition. However, it gives very good results in TESS dataset. Timbre is the feature of musical instrument, similarly, timbre in humans is a feature dependent on vocal cord and vocal tract. The length of vocal tract system in females is smaller and vocal folds are thinner (i.e., less mass) than males. TESS containing only female speakers and thus, affect the accuracy.

The combined feature with CQHC and CQPC contains both the pitch and the spectral components, i.e., comprises information regarding the harmonics, Fundamental frequency, and melody (which incorporates rhythm), giving very high accuracy in TESS and ESD parallel (Table 5.1).

Table 5.1: Classification Accuracy Results. After [87].

Features	**ESD Parallel**	**ESD Non-Parallel**	**TESS**
MFCC	96.63	36.05	81.34
GFCC	98.74	41.77	**99.87**
CQCC	95.83	61.08	79.31
CQT	96.34	57.82	82.74
CQHC	87.00	50.11	95.33
CQPC	**99.51**	50.51	**99.60**
CQHC + CQPC	**98.94**	**61.77**	**99.56**

5.5.3 Performance on Non-Parallel Dataset

Extending the analysis to non-parallel dataset, the major difference seen is the drastic decrease in the test accuracy across all features (Table 5.1), thus restablishing the model's dependence on speaker dependent information for classification.

The significant difference between Mel scale and CQT is that Mel scale follows a decadic logarithmic scale [78]. However, CQT follows a binary logarithmic scale, offering higher low-frequency resolution and better results (see Table 5.1). Eventhough, MFCC is time-invariant and is less susceptible to frequency shifts and more stable to time-warps, it is only upto a specific frame duration (20 ms), which is lesser than desired speech segments with lengths more than 250 ms needed to capture emotions better [78], thereby giving *36.05 %* accuracy.

The difference with time-invariance in CQT is that the time window is not fixed. This property also helps it to capture emotion-relevant information (pitch frequency) better. With higher invariance, irrelevant information like speaking style, speaker, etc. can be reduced, thus aiding in better classification. This property is extremely useful for emotions such as, anger and happiness (both carry higher energy content at higher frequencies) (Table 5.4). This difference is clearly seen in Table 5.1 as all CQT and CQT-based features perform better than baseline features MFCC and GFCC. The fusion feature CQHC+CQPC beats the baseline CQCC for speaker-independent emotion recognition by *0.69 %*. The vocal tract and vocal fold physiology and high resolution pitch information aid in this performance.

5.6 Statistical Measures

Experiments were performed to understand the test accuracy results. Statistical metrices, such as $F1$-score, MCC, Jaccard index and Hamming loss were performed for all the cepstral features obtained from ESD parallel and non-parallel data (Table 5.2 and 5.3). The values are rounded off at *0.2* decimal place. The quick variations in higher frequencies can be captured in CQT-based features due to better time resolution (for emotions, such as anger, and happiness). Table 5.4 gives the confusion matrix establishing the same.

Table 5.2: Performance Evaluation for Various Feature Sets on CNN Classifier on Parallel ESD. After [87].

Feature Set	**F1-Score**	**MCC**	**Jaccard Index**	**Hamming Loss**
MFCC	0.97	0.96	0.93	0.03
GFCC	0.99	0.98	0.98	0.01
CQCC	0.96	0.95	0.92	0.04
CQT	0.96	0.95	0.93	0.04
CQHC	0.87	0.84	0.77	0.13
CQPC	0.99	0.99	0.98	0.01
CQHC + CQPC	0.98	0.98	0.97	0.01

Table 5.3: Performance Evaluation for Various Feature Sets on CNN Classifier on Non-Parallel ESD. After [87].

Feature Set	**F1-Score**	**MCC**	**Jaccard Index**	**Hamming Loss**
MFCC	0.32	0.22	0.20	0.64
GFCC	0.37	0.30	0.24	0.58
CQCC	0.58	0.53	0.44	0.38
CQT	0.56	0.48	0.41	0.42
CQHC	0.47	0.38	0.34	0.50
CQPC	0.49	0.39	0.35	0.49
CQHC + CQPC	0.59	0.54	0.44	0.38

Table 5.4: Confusion Matrix Obtained for MFCC and CQHC+CQPC using CNN for Parallel ESD. After [87].

Feature	Emotions	Anger	Happy	Neutral	Sad	Surprise
MFCC	**Anger**	683	1	0	4	12
	Happy	13	644	9	3	31
	Neutral	3	1	688	7	1
	Sad	4	5	12	679	0
	Surprise	8	3	1	0	688
CQHC+CQPC	**Anger**	692	3	3	0	2
	Happy	6	684	5	0	5
	Neutral	0	1	698	1	0
	Sad	1	1	2	696	0
	Surprise	4	3	0	0	693

5.7 Chapter Summary

This study presented features motivated from the music literature for the challenging task of SER. Prosody in itself is an ever-developing concept, which is the key acoustic cue for emotion parameters. We captured three prosodic features, i.e., resolved pitch extraction (with CQPC), harmonic coefficients (with CQHC) which gives us fundamental tone and higher frequencies or upper harmonics and rhythm. The results illustrate that musical features work well for emotion recognition, as music stimulates emotion. With the results obtained, we find that acoustic patterns in frequency, energy, and spectral for the expression of different emotions are similar in speech and music, as mentioned in [74]. Proposed methodology (CQPC and CQHC+ CQPC) outperformed all baseline features expect GFCC for TESS dataset. The results prove that the proposed methods are extremely efficient for non-parallel (speaker-independent) classification. The next chapter proposes a new feature (LFRCC) for SER.

CHAPTER 6

Excitation Source-Based Features for SER

From the literature survey, it was observed that excitation source-based features also help in SER, but this field has yet to be explored. This chapter introduces a new feature for SER based on excitation source information of speech samples.

6.1 Introduction

Speech signals consist of both source (excitation source) and system (vocal tract) information. In the domain of Speech Emotion Recognition (SER), various features, including prosodic, source, and system-based features, have been explored. However, the exploration of excitation source information for SER is relatively limited [43]. To address this gap, the use of Linear Frequency Residual Cepstral Coefficients (LFRCC) is proposed in this study.

The effectiveness of LFRCC features has been demonstrated in the field of anti-spoofing [81]. To evaluate their performance in SER, a comparison is conducted against state-of-the-art features, namely MFCC and LFCC, using deep learning models such as ResNet and Time Delay Neural Network (TDNN) with Attention Statistics Pooling. The results indicate that the proposed LFRCC features outperform MFCC and LFCC by **25.64 %** and **10.25 %**, respectively, when utilizing the ResNet classifier. Similarly, when employing the TDNN classifier, the proposed features achieve a performance improvement of **12.82 %** and **5.31 %** over MFCC and LFCC, respectively. These findings highlight the efficacy of the proposed LFRCC features for SER compared to the commonly used MFCC and LFCC features. Further, classifier-level and score-level fusion were performed, and MFCC+LFRCC gave the highest accuracy of **92.31 %**. The importance of context and the relevance of respiration patterns in SER is studied.

6.2 Linear Prediction (LP) Residual

The LP method has a historical foundation in speech coding applications derived from the literature on system identification and control [51]. In LP analysis, each speech sample is expressed as a linear weighted combination of preceding ′p′ speech samples, with ′p′ denoting the order of the linear predictors. The coefficients assigned to these weights are referred to as Linear Prediction Coefficients (LPCs) [51]. Specifically, if s(n) represents the current speech sample, the predicted sample can be expressed as follows:

$$\hat{s}(n) = -\sum_{k=1}^{p} a_k s(n-k), \tag{6.1}$$

where a_k are LPCs. The difference between the actual speech sample $s(n)$ and the predicted samples $\hat{s}(n)$ is known as LP residual i.e., $r(n)$ and is given by:

$$r(n) = s(n) - \hat{s}(n) = s(n) + \sum_{k=1}^{p} a_k s(n-k). \tag{6.2}$$

In particular, if we take the all-pole inverse filtering of the speech signal using the above LP analysis, we have

$$A(z) = 1 + \sum_{k=1}^{p} a_k z^{-k} \tag{6.3}$$

$$H(z) = \frac{G}{1 + \sum_{k=1}^{p} a_k z^{-k}}. \tag{6.4}$$

The equation provided represents the predicted sample in terms of the inverse filter $A(z)$, which corresponds to the all-pole LP filter $H(z)$ capturing the characteristics of the vocal tract-based system. Additionally, the term G refers to the gain in the LP model. To enhance the system information with excitation source information, the LP residual is utilized. The LP residual captures the characteristics of the excitation source signal. To extract and represent this information, the LP residual is processed in the cepstral domain, allowing for the representation of the spectral envelope of the excitation source signal.

6.3 Linear Frequency Residual Cepstral Coefficients (LFRCC)

Figure 6.1 illustrates the functional block diagram of the proposed feature set. The input speech signal undergoes pre-emphasis filtering to balance the lower and higher frequency components [81]. Subsequently, the signal is processed by the LP block, resulting in the LP residual waveform, denoted as $r(n)$. The LP residual waveform is then divided into frames and subjected to windowing with a duration of 25 ms and a frame shift of 15 ms. In the next step, the power spectrum is estimated for each frame of the LP residual and passed through a filterbank consisting of 40 linearly-spaced triangular subband filters. To obtain the desired LFRCC features with minimal distortion, the Discrete Cosine Transform (DCT) and Cepstral Mean Normalization (CMN) techniques are applied to the power spectrum. This sequence of operations yields the final LFRCC features that capture relevant characteristics of the speech signal.

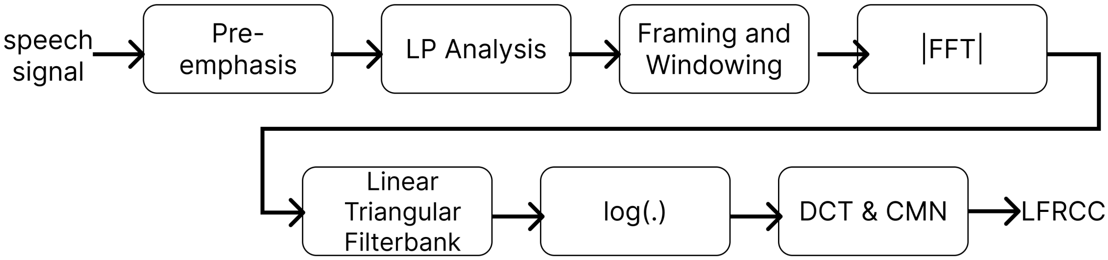


Figure 6.1: Schematic block diagram of LFRCC feature extraction. After [81].

6.4 Experimental Details

The state-of-the-art MFCC and LFCC features are used for comparing the proposed features. *39*-D coefficients were taken containing static, delta, and double-delta parameters. The window length taken is 25 ms, the number of subband filters used is 40, and the number of points used in the Fast Fourier Transform (NFFT) is 512. EmoDB dataset is used for experimentation. The current investigation focuses on four emotions, anger, happiness, neutrality, and sadness, with one male speaker reserved for the test. Classifier details are given in Chapter 3.

6.5 Spectrographic Analysis

Figure 6.2 and Figure 6.3 present the spectrograms of the original signal and the LP residual signal for both a female and a male speaker uttering the same sentence. In comparison to the spectrogram of the original signal, the LP residual spectrogram exhibits distinct characteristics such as the clear representation of formants, harmonics (horizontal pitch striations), and energy distribution. A notable observation shared across all emotions is the prominent energy presence at higher frequencies in the LP residual spectrograms, as highlighted by the black box in Figure 6.2 and Figure 6.3. Moreover, the width between the horizontal striations is found to be greater for females due to their higher fundamental frequency (F_0) compared to males. Notably, anger exhibits short pauses caused by irregular breathing (puffs) in contrast to neutral and sad emotions, which feature longer pauses and low formant fluctuation due to deeper breathing. A significant distinction between happy and anger emotions lies in the distribution of high energy content. In the case of happy emotions, the high energy density is evenly spread throughout the utterance, gradually diminishing towards the end. Conversely, anger emotions concentrate the high energy density at higher frequencies, maintaining it consistently throughout the utterance, as clearly depicted in the LP residual spectrogram. These observations were made by analyzing multiple sentences, one of which is represented in Figure 6.2 and Figure 6.3.

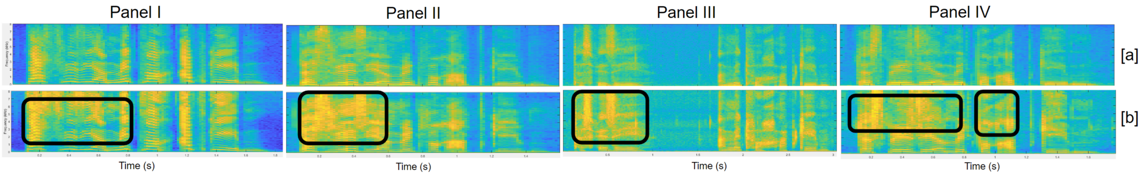


Figure 6.2: Spectrographic analysis for (a) original speech signal and (b) corresponding LP residual. Panel I(a), Panel II(a), Panel III(a), and Panel IV(a) represent the spectrograms of a female speaker for the emotions anger, happy, sad, and neutral, respectively, for the sentence *"Das will sie am Mittwoch abgeben (She will hand it in on Wednesday)".*

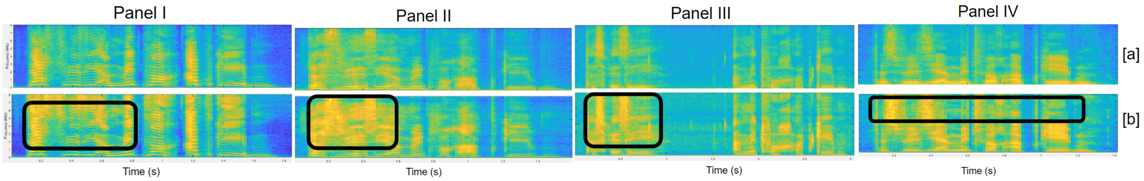


Figure 6.3: Spectrographic analysis for (a) original speech signal and (b)corresponding LP residual. Panel I(a), Panel II(a), Panel III(a), and Panel IV(a) represent the spectrograms of a male speaker for the emotions anger, happy, sad, and neutral, respectively, for the sentence *"Das will sie am Mittwoch abgeben (She will hand it in on Wednesday)".*

6.6 Effect of LP Order

In the proposed method, the LFRCC feature set is obtained by varying the prediction order (p) from 4 to 25 for a sampling frequency of *16*KHz, as depicted in

Figure 6.4. The results indicate that the highest classification accuracy is achieved when using an LP order of *20*. Remarkably, this optimal order remains consistent regardless of the classifier employed. For both TDNN and ResNet, an LP order of *20* yields the highest accuracy rates of **89.74 %** and **87.17 %**, respectively (as illustrated in Figure 6.4). Additionally, it is worth noting that the accuracy of emotion classification tends to be higher for higher LP orders (16-25) compared to lower LP orders (4-15). This observation can be attributed to the fact that higher LP orders allow for better capture of contextual information of emotional aspects, especially concerning speech prosody, which is predominantly *suprasegmental* in nature and thus, requires a longer duration of speech signal.

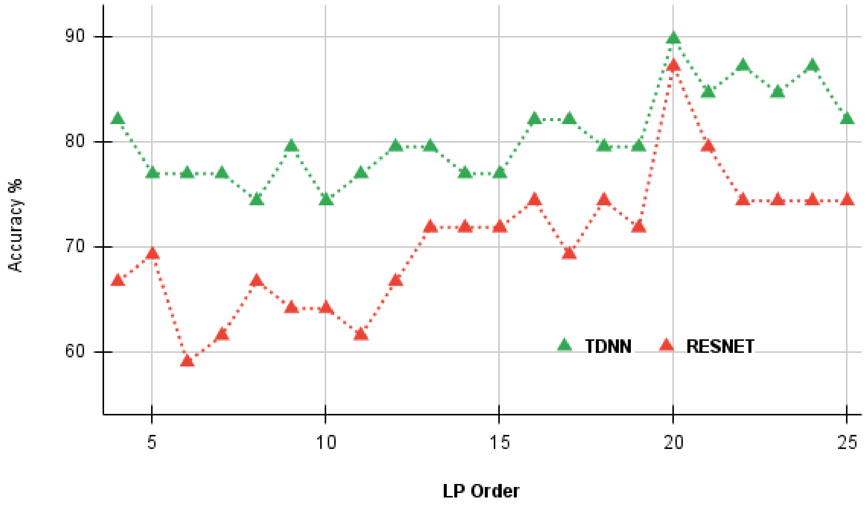


Figure 6.4: Effect of LP orders for LFRCC using TDNN and ResNet Classifiers. After [86].

6.7 Results with Score-Level Fusion

To understand the complementary information captured by different features, score-level fusion is performed using the following data fusion strategy.

$$L_{fused} = \alpha L_{feature1} + (1-\alpha) L_{feature2}, \tag{6.5}$$

where $L_{feature1}$ is a raw score of either MFCC or LFCC, whereas $L_{feature2}$ represents the raw score of LFRCC. Figure 6.5 and Figure 6.6 depicts the score-level fusion on TDNN and ResNet, respectively. It is observed that MFCC+LFRCC gives the best classification accuracy of **92.31 %** and *84.62 %* in TDNN and ResNet classifier, respectively. LFCC+LFRCC performs better with ResNet than MFCC +LFRCC by **2.56 %**, whereas with TDNN, it gives 89.74%. MFCC+LFCC does not show any significant improvement. Based on the obtained results, it can be concluded that score-level fusion with TDNN outperforms ResNet. This difference in performance can be attributed to the temporal dependency modeling capability of TDNN, which allows it to capture local patterns in the data. In contrast, ResNets are more suitable for capturing global patterns in images, such as spatial relationships between objects. When considering the combination of features, MFCC+LFRCC achieves the highest classification results. This is because MFCC effectively captures spectral information in the lower frequency regions of speech, while LFRCC captures excitation information in the higher frequency regions. By combining these two feature sets, the major emotional context in speech signals can be captured comprehensively.

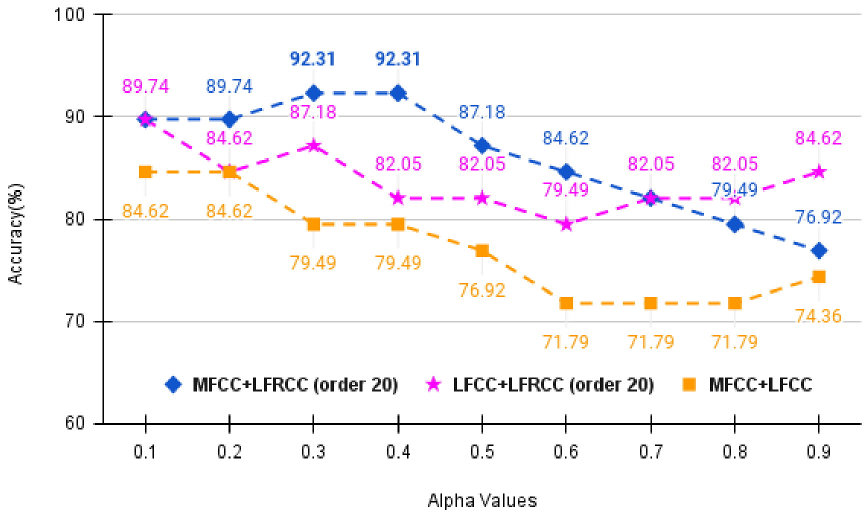


Figure 6.5: Score-level fusion of features using TDNN classifier. After [86].

6.8 Results with Classifier-Level Fusion

Figure 6.7 depicts the classifier-level fusion, i.e., the output of each classifier (TDNN and ResNet) is multiplied by a weight. Then all the weighted outputs are added together to obtain the final output. It is observed that the highest classification accuracy obtained for MFCC is **76.92 %**, LFCC is **84.62 %**, and LFRCC is **89.74 %**, which are the results for α= 1, which implies results from TDNN classifier alone. This indicates that the classifiers' characteristics together could not add enough weight to the accuracy, which we are getting from features (MFCC+LFRCC). This

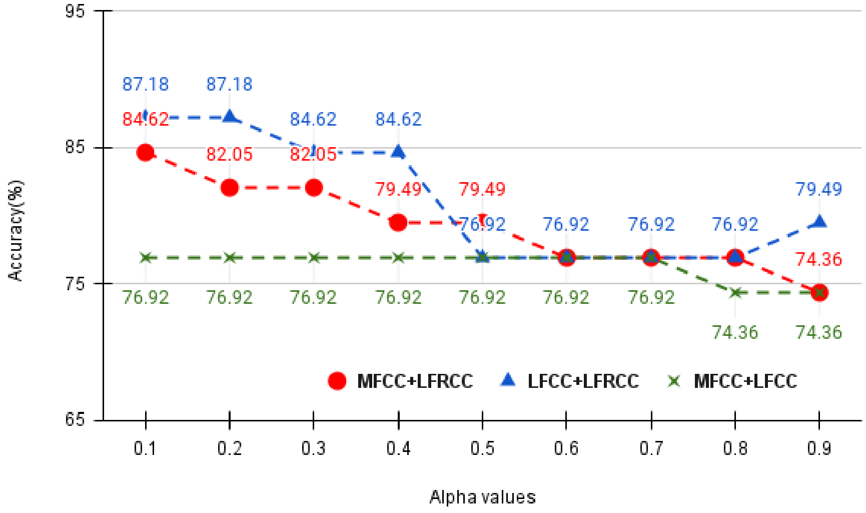


Figure 6.6: Score-level fusion of features using ResNet classifier. After [86].

proves that the results obtained are not significantly dependent on the classifier, thereby proving that the proposed LFRCC feature set captures emotion information well.

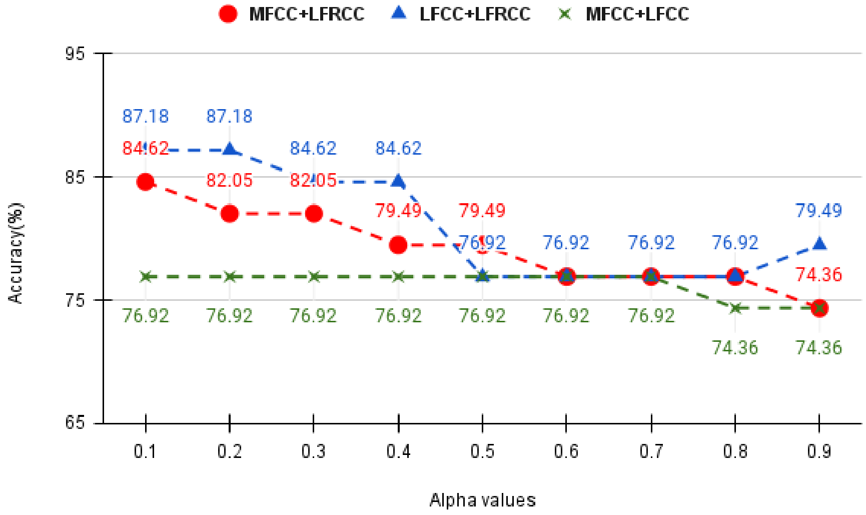


Figure 6.7: Classifier-level fusion for single feature sets using TDNN and ResNet classifiers. After [86].

6.9 Performance of LFRCC on SER

From Table 6.1, it is observed that the cepstral coefficients with linear filterbanks (i.e., LFCC) results in better classification than the Mel filterbanks (i.e., MFCC) irrespective of the classifiers used for SER. This is because, LFCC captures emotion information well in higher frequency regions as compared to the MFCC as the width of the triangular filters in MFCC increases with frequency and thus, ignoring fine spectrl details. The emotions, in particular, anger and happy operate in higher frequency regions (Section 6.5), which is captured better by LFCC due to constant difference between the width of subband filters in filterbank throughout.

The LFRCC feature set demonstrates superior performance compared to the baseline MFCC and LFCC features, achieving a **25.64 %** and **10.25 %** improvement, respectively, in ResNet, and a **12.82 %** and **5.31 %** improvement, respectively, in TDNN (as shown in Table 6.1). It is widely acknowledged that the lungs play a crucial role in providing the necessary airflow, which acts as a power supply for speech production. Consequently, changes in respiratory patterns directly impact the timing, duration, and overall rhythm of speech. As a result, respiratory patterns can influence the emotional expression in speech signals[37], [34]. The LFRCC feature set effectively captures this excitation source information (as discussed in Section 6.5). Specifically, the section around the Glottal Closure Instants (GCI) in the LP Residual signal exhibits a high signal-to-noise ratio (SNR) due to the impulse-like excitation, and during the Glottal Closure (GC) phase, the excitation source is completely isolated from the vocal tract [44]. This region contains valuable information that cannot be adequately captured by MFCC and LFCC, making LFRCC an effective choice for representing these features. Though in isolation, the Mean Square Error (MSE) is not zero, but a very low value and this value helps capture the formant information from LFRCC. It is also observed that the pitch information from LP residual audio was void of any experimental noise (such as that of mic, etc) and thus, gave pure speech, indicating its noise robustness attribute and high classification accuracy.

Table 6.1: Classification Accuracy (in %). After [86].

Classifier	**MFCC**	**LFCC**	**LFRCC**
ResNet	61.53	76.92	**87.17**
TDNN	76.92	84.61	**89.74**

6.10 Chapter Summary

In this study, a novel feature set for emotion recognition, called excitation source-based LFRCC, was introduced. To assess its performance, traditional vocal tract system features such as MFCC and LFCC were also used for comparison. The aim was to leverage the additional information provided by the LP residual-based feature, which demonstrated superior performance compared to existing spectral features. This finding suggests that the proposed features possess discriminative power for classifying emotions. Additionally, the study observed the significance of linear filterbank in contrast to the Mel filterbank. Furthermore, the results obtained from classifier-level fusion highlighted the effectiveness of the proposed feature set, irrespective of the specific classifier employed. In the subsequent chapter, the research delves into the exploration of phase-based features to extract vocal fold and vocal state information for the task of SER .

CHAPTER 7

Vocal Tract-Based Features for SER

After analyzing prosodic and excitation source features, this chapter proposes a new phase-based feature that captures vocal tract system information for SER.

7.1 Introduction

Prosodic features, such as pitch, fundamental or pitch frequency (F_0), duration, energy, and others, are widely employed for SER in the literature [80]. Nonetheless, these features are limited to characterizing only the vocal folds state. Therefore, incorporating a feature that characterizes both the vocal tract and vocal fold state would enhance the emotion classification performance.

This study investigates the utility of phase spectrum on SER as it performs well in speech recognition [75] and in the extraction of source and system information [54]. To extract fine structures from the spectral envelope, we employ group delay and modified group delay functions, as they are found to capture system information more effectively than the magnitude or Linear Prediction (LP) spectrum. The robustness of these proposed features is also tested with state-of-the-art features used for SER. To our knowledge, this work is the first attempt to use the MGDCC feature on emotions. Experiments were performed using the EmoDB database on emotions, anger, happy, neutral, and sad. The proposed feature outperformed the baseline Mel Frequency Cepstral Coefficients (MFCC) and Linear Frequency Cepstral Coefficients (LFCC) by **7.7 %** and **5.14 %**, respectively. The noise robustness characteristics of MGDCC were tested for stationary and non-stationary noise, and the results were promising. The latency period was also analyzed and MGDCC proved to be the most practically suitable feature.

7.2 Phase-Based Features

Extracting phase-based features is challenging since the frequency domain's phase spectrum is discontinuous. For the phase to be used, it has to be unwrapped to make it a continuous function. However, the phase unwrapping technique is computationally complex due to its non-uniqueness. On the other hand, the group delay and modified group delay techniques have similar properties to the unwrapped phase and are known to be extracted directly from the signal [55].

7.2.1 Importance of Phase Information in Emotions

The phase information in speech pertains to the speech signal's temporal properties, encompassing the waveform's alterations as time progresses. Within SER, the phase information assumes a prominent position and offers invaluable perspectives into the emotional essence conveyed by speech. Presented below are several justifications elucidating the importance of phase information in speech emotion recognition:

- **Prosodic Features**: Emotional expressions in speech encompass the spectral aspects and prosodic features like pitch, intonation, and rhythm. The phase information is closely linked to these prosodic characteristics, and its utilization can significantly contribute to the precise detection and analysis of emotional cues. By examining variations in the phase, researchers can capture changes in pitch contours, temporal dynamics, and expressive patterns that convey specific emotions.

- **Speech Dynamics**: Emotional speech often exhibits discernible temporal variations and dynamic patterns. The phase information can furnish additional insights into speech's temporal structure and rhythmic properties, enabling the recognition of emotions characterized by distinct speech dynamics. By incorporating the phase information, models for SER can effectively capture the subtle fluctuations in speech dynamics associated with different emotional states.

- **Emotion-specific Articulation**: Certain emotions are known to impact speech articulation and the configuration of the vocal tract. The phase information carries crucial details about the alignment and synchronization of various speech components, such as glottal excitation and vocal tract resonances.

Analyzing phase cues can facilitate identifying emotion-related changes in articulatory patterns that cannot be solely captured by spectral information.

- **Perception of Emotional Cues**: Human listeners rely on spectral and temporal acoustic cues to perceive and interpret emotions in speech. The phase information contributes to the perceptual experience of speech and can influence the processing and comprehension of emotional cues. By incorporating phase information, SER models can align with the human perception of emotional speech, enhancing their ability to recognize and interpret emotions accurately.

- **Robustness to Noise and Degraded Speech**: Using phase information can enhance the robustness of SER models in adverse acoustic conditions, including background noise or degraded speech signals. While spectral information is susceptible to degradation, phase information can provide valuable cues for emotion recognition even when the spectral content is compromised. Considering both spectral and phase features can improve the reliability and generalizability of SER systems, particularly in challenging and real-world scenarios.

7.2.2 Group Delay and Modified Group Delay Function

The group delay function can be obtained by taking the negative derivative of the phase of the unwrapped Fourier transform. Additionally, the group delay of the signal p(n) can be calculated using the following method based on the signal itself:

$$T_m(\omega) = -Im\frac{d(P(\omega))}{d\omega}, \tag{7.1}$$

upon solving the Eq (7.1) as stated in [55], we arrive at:

$$T_m(\omega) = \frac{P_R(\omega)Q_R(\omega) + P_I(\omega)Q_I(\omega)}{|P(\omega)|^2}. \tag{7.2}$$

where $P(\omega)$ and $Q(\omega)$ are Fourier transforms of $p(n)$ and $np(n)$, respectively. The $P_R(\omega)$ and $P_I(\omega)$ indicates the real and imaginary parts of $p(\omega)$, respectively. The group delay function in terms of cepstral coefficients can be expressed as [33]:

$$T_m(\omega) = \sum_{n=1}^{+\infty} nc(n)cos(n\omega), \tag{7.3}$$

where $c(n)$ indicates the n-dimensional cepstral coefficients. This operation is replicated by applying Discrete Cosine Transform (DCT). The two most important properties of group delay feature that gives them an edge compared to magnitude-based features are **additivity** and **high resolution**. Nevertheless, despite the advantages mentioned, the effective utilization of the group delay function in speech processing tasks relies on the signal being a *minimum phase*. In the case of a non-minimum phase signal, the presence of roots of the Z-transformed signal outside or near the unit circle leads to spikes in the group delay spectrum. These spikes cause distortions in the fine structure of the vocal tract system's envelope and mask the location of formants [33]. The occurrence of these spikes is a result of a smaller denominator term in Eq (7.2), indicating a smaller distance between the corresponding zero location and the frequency bin on the unit circle.

Any meaningful use of the phase-based features comes with the reduction of inadvertent spikes due to the smaller denominator value in Eq (7.2). One such representation is the **modified group delay function (MODGF)** introduced to maintain the dynamic range of the group delay spectrum. The MODGF is given by [33]:

$$T_m(\omega) = \frac{T(\omega)}{|T(\omega)|}|T(\omega)|^{\alpha}, \tag{7.4}$$

where

$$T(\omega) = \frac{P_R(\omega)Q_R(\omega) + P_I(\omega)Q_I(\omega)}{|S(\omega)|^{2\gamma}}, \tag{7.5}$$

where $S(\omega)$ represents the cepstrally-smoothed version of $|P(\omega)|$. It was seen that introducing $|S(\omega)|$, very low values can be avoided. The parameters α and γ are introduced to reduce the spikes and restore the dynamic of the speech spectrum, respectively. Both parameters α and γ vary from 0 to 1. DCT is applied to convert the spectrum to cepstral features to obtain the cepstral coefficients. The first coefficient of the cepstral coefficients is ignored as this value corresponds to the average value in the GDF.

7.2.3 Robustness of Modified Group Delay Function

In this section, we analytically show the robustness of the group delay function to additive noise, which also applies to the modified group delay function [61]. Let $u(n)$ denote a clean speech signal, degraded by adding uncorrelated, additive noise $v(n)$ with 0 mean and σ^2 variance. Then, the noisy speech $z(n)$ can be expressed as,

$$z(n) = u(n) + v(n). \tag{7.6}$$

Taking the Fourier transform and obtaining the power spectrum we have,

$$P_z(\omega) = P_u(\omega) + P_v(\omega). \tag{7.7}$$

From eq (7.7), there can be two mutually exclusive frequency regions, namely, *high SNR* and *low SNR*.

Low SNR

Considering a low SNR situation, i.e., $P_u(\omega) << \sigma^2(\omega)$ (where noise power is $\sigma^2(\omega)$ due to the assumption that noise is having 0 mean), we have:

$$P_z(\omega) = \sigma^2(\omega)(1 + \frac{P_u(\omega)}{\sigma^2(\omega)}). \tag{7.8}$$

Taking the logarithm on both sides and using the Taylor series expansion and ignoring the higher order terms results in:

$$ln(P_z(\omega)) \approx ln(\sigma^2(\omega)) + \frac{P_u(\omega)}{\sigma^2(\omega)}. \tag{7.9}$$

Since $P_u(\omega)$ is a continuous and periodic function of ω, it can be expanded using the Fourier series. In particular,

$$ln(P_z(\omega)) \approx ln(\sigma^2(\omega)) + \frac{1}{\sigma^2(\omega)}\left[\frac{d_0}{2} + \sum_{k=1}^{+\infty} d_k cos(\frac{2\pi}{\omega_0}\omega k)\right], \tag{7.10}$$

where d_k's are the Fourier series coefficients. Since $P_u(w)$ is a power spectrum, it is an even function, and the coefficients of sine terms are zeros. To relate the spectral phase and magnitude with the cepstral coefficients, let us consider the Fourier transform representation of a sequence b(n):

$$B(e^{j\omega}) = |B(e^{j\omega})|e^{j\theta}(e^{j\omega}). \tag{7.11}$$

Since the log-magnitude component is an even function, the resulting Fourier series expansion can be given by:

$$ln(|B(e^{j\omega})|) = \frac{c[0]}{2} + \sum_{n=1}^{+\infty} c[n] cos(\omega n). \tag{7.12}$$

From the properties of the Fourier phase spectrum, the phase spectrum is an odd function. Hence the resulting Fourier series expansion is given by:

$$\theta(e^{j\omega}) = -\sum_{n=1}^{+\infty} c[n]sin(\omega n), \tag{7.13}$$

where c[n] are the cepstral coefficients. Group delay coefficients are obtained by considering the negative logarithm of the unwrapped phase obtained in eq (7.13):

$$T(e^{jw}) = -\sum_{n=1}^{+\infty} nc[n]cos(\omega n). \tag{7.14}$$

From eq's (7.12) and (7.13), it can be observed that the phase and log magnitude spectra of a signal are related through the cepstral coefficients. Assuming the additive noise as a minimum phase signal [33]. From eq's (7.12), (7.13), and (7.14), it can be observed that the group delay function can be extracted from the log magnitude response by ignoring the DC term and multiplying each coefficient by k. Applying this observation to eq (7.10) we can obtain the group delay function as [61]:

$$T_z(\omega) \approx \frac{1}{\sigma^2(\omega)} \sum_{k=1}^{+\infty} kd_k cos(\omega k). \tag{7.15}$$

Eq (7.15) indicates that the group delay function is inversely proportional to the noise power in the regions with low SNR. This indicates that the group delay function preserves peaks and valleys well in the presence of additive noise.

High SNR

Now consider frequencies such that $P_u(\omega) >> \sigma^2(\omega)$, we have:

$$P_z(\omega) = P_u(\omega)(1 + \frac{\sigma^2(\omega)}{P_u(\omega)}). \tag{7.16}$$

Taking the logarithm on both the sides and using the Taylor series expansion results in:

$$ln(P_z(\omega)) \approx ln(P_u(\omega)) + \frac{\sigma^2(\omega)}{P_u(\omega)}. \tag{7.17}$$

Since $P_u(w)$ is a non-zero, continuous, periodic function of ω, the same can be said about $\frac{1}{P_u(\omega)}$. Hence both ln(.) and $(\frac{1}{.})$ of eq (7.17) can be expanded using the

Fourier series, resulting in:

$$ln(P_z(\omega)) \approx \frac{d_0}{2} + \frac{\sigma^2(\omega)e_0}{2} + \sum_{k=1}^{+\infty}(d_k + \sigma^2(\omega)e_k)cos(\omega k), \tag{7.18}$$

where d_k's and e_k are the Fourier series coefficients of $ln(P_u(\omega)$ and $\frac{1}{P_u(\omega)}$ respectively.

Using eq's (7.12) and (7.13), we obtain the group delay function as:

$$T_z(\omega) \approx \sum_{k=1}^{+\infty} k(d_k + \sigma^2(\omega)e_k)cos(\omega k). \tag{7.19}$$

Eq (7.19) indicates that the noise power $(\sigma^2(\omega))$ is negligible when the signal power is higher than the noise power, and the group delay function can be expressed only using the log-magnitude spectrum. This conveys that the group delay spectrum follows the signal's envelope rather than noise. Hence, it preserves the formant peaks well in the presence of additive noise [61].

7.3 Experimental Details

The state-of-the-art features, MFCC and LFCC, are used for comparison. To maintain uniformity among features, *20*-D feature vectors with a window length of *25*ms and a hop length of *10*ms are used for all. The details of the classifier is discussed in Chapter 3. To restore the dynamic range of phase-based features, MGDCC has two additional constraint parameters, alpha (α) and gamma (γ). The CNN classifier is used to fine-tune these parameters by varying them from 0 to 1 with a step size of *0.1*. The optimal parameters thus found by classification accuracy is **α=0.1, γ=0.1** (Refer Figure 7.1).

7.4 Experimental Results

7.4.1 Spectrographic Analysis

Panel-A, Panel-B, and Panel-C of Figure 7.2 represent the Spectrogram, Mel Spectrogram, and MGDCC-gram analysis of various emotions, respectively. Figure 7.2(a), Figure 7.2(b), Figure 7.2(c), and Figure 7.2(d) show the analysis for anger, happy, sad, and neutral, respectively. Mel spectrograms give a broad and dull representation of utterance and thus have obstructed the ability to identify the fine

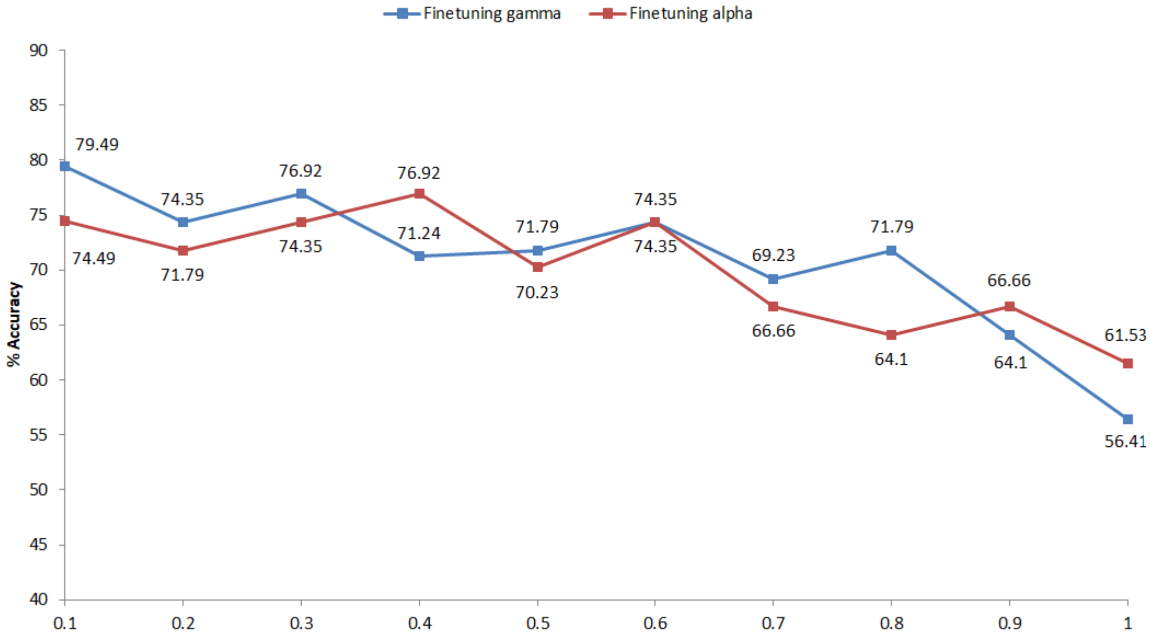


Figure 7.1: Tuning Parameters α and γ using Greedy Search Technique for Emotion Recognition. After [85].

formant structures and energy distribution. It can be observed from the plots that the fine structure of the formants that can be observed in the magnitude spectrum (Panel-A) can also be seen in the spectrogram obtained by the modified group delay spectrum. Hence, there is no information loss while using phase-based cepstral coefficients. Additionally, the resolution between the formants is high in the phase-based, i.e., modified group delay spectrum resulting in a better distinction among the formants. This is due to the fact that the denominator term at the formant frequencies becomes 0 (as the pole radius approaches to unit circle) resulting in peaks that give a higher resolution formants. Additionally, phase features are able to capture irregularities in the speech signal. The presence of turbulence in a speech signal changes with emotion and these irregularities are captured better through phase signal rather than the magnitude spectrum.

7.4.2 Comparison with Baseline Features

From Table 7.1, it can be observed that LFCC is the best-performing baseline feature. The MGDCC feature outperforms the magnitude-based features i.e., MFCC and LFCC by a margin of **7.7**% and **5.14** %, respectively. This might be because of the high-resolution property of the modified group delay function which can be noticed in Figure 7.2. They capture the fine structures of spectral envelope and thus formant structures are emphasized well. However, GDCC fails to achieve

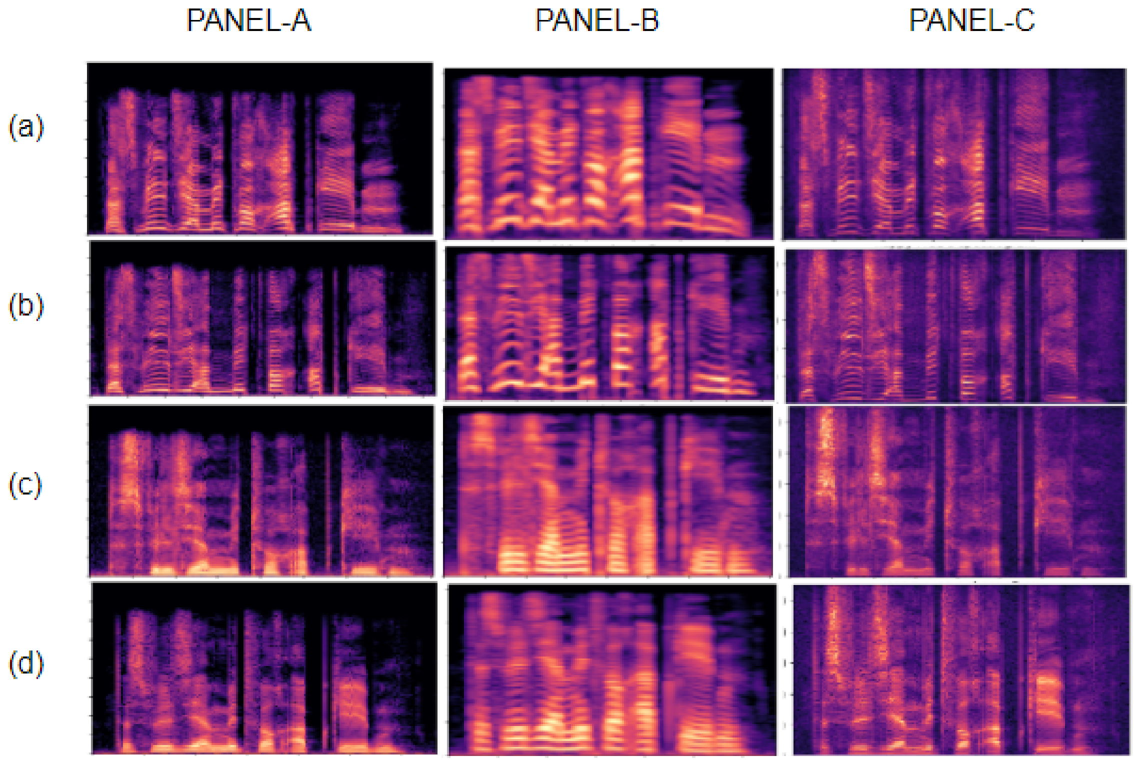


Figure 7.2: Panel-A, Panel-B, and Panel-C represent the Spectrograms *vs.* Mel Spectrograms *vs.* MGD spectrogram of a male speaker uttering the same sentence in emotions- (a) anger, (b) happy, (c) sad, and (d) neutral, respectively. After [85].

similar performance. This is because of the noisy structure resulting from the GDCC occurring from the presence of zeros close to or outside the unit circle. These spikes cause formant masking, making it difficult to obtain valuable features for the classification task. It is also observed LFCC captures emotional information well in higher frequency regions compared to MFCCs as in MFCC, the width of the triangular filters increases with frequency and thus, ignoring fine details (Section 7.4.1). The emotions, in particular, anger and happy operate in higher frequency regions, which is captured better by LFCC due to the constant difference between the width of filterbanks throughout.

Table 7.1: Classification Accuracy on CNN. After [85].

Feature Set	**MFCC**	**LFCC**	**GDCC**	**MGDCC**
Test Acccuracy	71.79	74.35	56.41	**79.49**

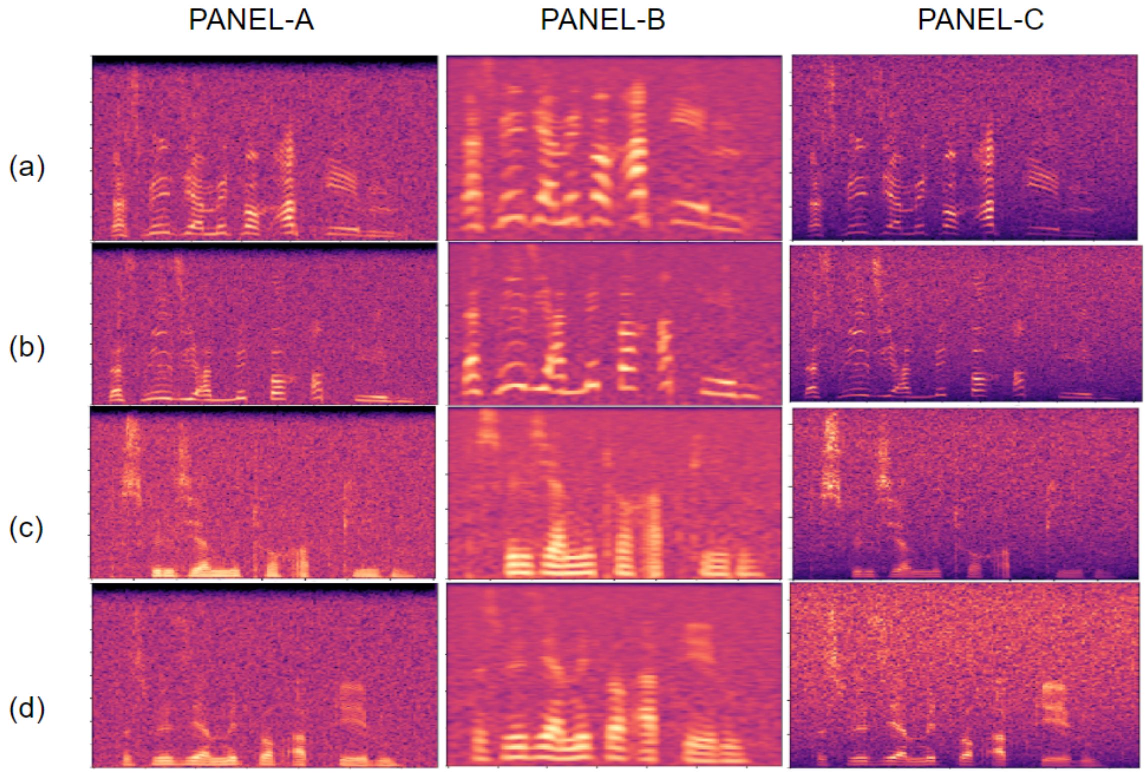


Figure 7.3: Panel-A, Panel-B, and Panel-C represent the Spectrograms *vs.* Mel Spectrograms *vs.* MGD spectrogram of white noise added speech of a male speaker uttering the same sentence in emotions- (a) anger, (b) happy, (c) sad, and (d) neutal, respectively. After [85].

7.4.3 Performance of Proposed Feature under Signal Degradation

The robustness of the proposed features is tested using various noise types, such as white, pink, babble, and street noise with SNR levels of -10 dB, -5 dB, 0 dB, 5 dB, 10 dB, and 15 dB. When we consider additive white noise for evaluation, due to the nature of AWGN, the noise is distributed across all the bands of frequency. From Table 2, at the low SNR values, MGDCC clearly outperforms both magnitude-based features, MFCC and LFCC by a significant margin of **3.41 %**, **10.25 %**, respectively. Similarly, at higher SNR values, MGDCC outperforms baseline features MFCC and LFCC by **17.95 %**, **7.79 %**, respectively. Considering that the signal is degraded by the pink noise, which has higher noise power in lower frequencies rather than the higher frequencies, the MGDCC feature set outperforms both MFCC and LFCC features. Additionally, when considered non-stationary noises (noises which vary w.r.t time) such as street noise or traffic noise and babble noise are considered. The MGDCC noise robustness is evident in any

Table 7.2: Classification Accuracy on CNN with different noise types on EmoDB. After [85].

NOISE	FEATURE	-10dB	-5dB	0dB	5dB	10dB	15dB
Babble	MGDCC	79.48	81.29	82.05	81.66	81.66	82.66
	MFCC	74.35	76	79.48	79.48	79.48	79.48
	LFCC	61.53	66.66	79.48	76.92	79.48	79.48
Street	MGDCC	75.35	80	81.66	81.66	71.79	86.66
	MFCC	74.35	76	76.92	79.48	88.48	82.05
	LFCC	74.35	70.23	79.48	71.79	76.92	79.48
White	MGDCC	76.92	79.48	74.35	71.79	76.92	74.35
	MFCC	69.23	76.92	74.35	43.58	82.05	43.58
	LFCC	71.79	64.10	64.10	58.97	71.79	69.23
Pink	MGDCC	74.35	69.23	71.79	71.79	71.79	74.35
	MFCC	41.79	38.46	41.02	43.58	70.35	71.79
	LFCC	71.79	66.66	66.66	61.53	71.79	71.79

kind of noise. These results indicate that the performance of the baseline features is degraded in the presence of stationary and non-stationary noise, whereas the performance of MGDCC remains intact across various noise types. These results prove the additive noise robustness property, and also that the group delay spectrum is known to emphasize the signal spectrum rather than the noise spectrum. It can also be explained by the fact that the MGDCC feature set pushes the zeros into the unit circle in an attempt of making the signal a minimum phase, which may also help in the suppression of noise. Additionally, it can be noted that LFCC and MFCC are not equally robust in white noise as the energy in higher frequency speech regions is weak making it more susceptible to noise corruption. The LFCC contains more subband filters at higher frequencies than MFCC, making it less robust to white noise. As the noise power decreases, the LFCC feature set still outperforms MFCC due to its linearly-spaced subband filters instead of the Mel filterbank. This reasoning also explains the comparable performance of MFCC to LFCC, when the signal is corrupted with pink noise.

7.4.4 Analysis of Latency Period

In this study, we investigated the latency period of the MGDCC feature set in comparison to the baseline features, i.e., MFCC and LFCC. To evaluate the performance of CNN based on different feature sets, we measured the accuracy % with respect to the latency period, as depicted in Figure 7.4. The latency period denotes the time elapsed between the utterance of speech and the system's response, expressed as a percentage fold accuracy that represents the number of frames utilized for utterance classification. Therefore, if the system demonstrates superior performance at lower latency periods, it implies that it can classify the speech utterance effectively without requiring a prolonged duration of speech. The duration of utterance is upto *3* sec and is plotted at an interval of *0.5* sec. It is observed that MGDCC features give significant classification performance throughout, the highest accuracy being *79.48* % at 1.5 sec (Figure 7.4). On the contrary, the baseline features constantly down-perform and take a longer duration to achieve comparable performance. This encourages the practical suitability of the proposed MGDCC feature set.

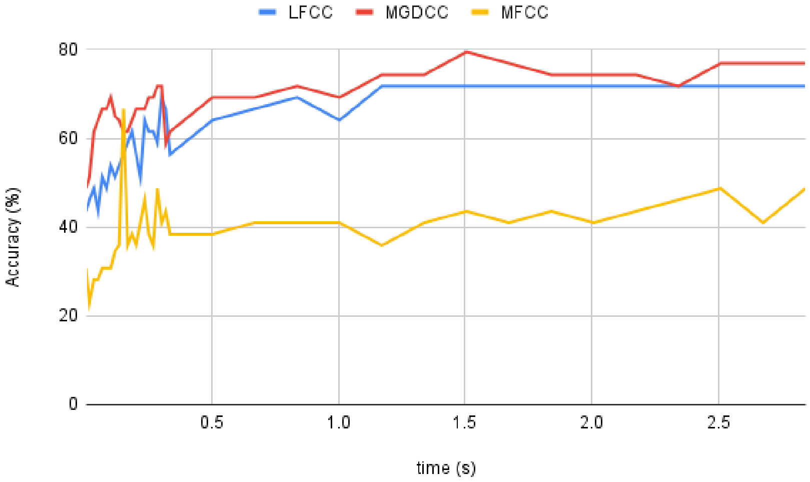


Figure 7.4: Latency Period of MFCC, LFCC, and MGDCC. After [85].

7.5 Chapter Summary

In this study, phase-based vocal tract features were proposed for emotion recognition. Other state-of-the-art spectral features MFCC and LFCC were used for comparison. The objective was to capture the irregularities in speech signal and the formant structure better for efficient SER. MGDCC also proved to perform well for stationary and non-stationary noise added dataset due to its additive noise robustness property. The significance of linear filterbanks over Mel filterbanks were observed for emotion classification. The practical suitability of MGDCC was also calculated and promising results were seen.

CHAPTER 8
Whisper Features for SER

8.1 Introduction

In human communication, speech is the most natural mode of expression and is inherently infused with emotions. Speech allows for greater emotional expression and understanding than the other modes, such as text messages or emails. Recognizing emotions in speech is crucial for improving the intractability between humans and machines. To develop an effective system for SER, a comprehensive understanding of the acoustics associated with different emotions is crucial.

In academic research, a neural network dubbed Whisper (Web-scale Supervised Pretraining for Speech Recognition) has been developed and released by Open AI [6]. This open source Automatic Speech Recognition (ASR) system is founded upon an expansive dataset comprising 680,000 hours of English-language audio data labeled via web collection, 117,000 hours of data for 96 other languages, and 125,000 hours of translational data. The usage of such a voluminous dataset has enabled Whisper to attain remarkable levels of precision, approximating human accuracy for speech recognition in English. The model's creators postulate that it can capture all crucial information from the speech signal and map it to a fixed-size vector upon the conclusion of its first layer, which can be utilized to execute SER task. The researchers contend that the effectiveness of these features is primarily attributed to the vast amount of data utilized to train the Whisper model, which enabled it to capture not just linguistic information but also acoustic and other pertinent details.

In this study, the performance of different feature extraction techniques (MFCC, GFCC, and Whisper) on three different databases, namely, ESD, CREMA, and EmoDB is done and it is found that Whisper features outperforms other feature extraction techniques (by **2-30 %**) in all scenarios. The study also demonstrates the feasibility of using Whisper features in cross-database scenarios, which has important implications for real-world applications. Overall, the results of this study suggest that Whisper features are a robust and effective approach to SER task.

To the best of our knowledge, this study is the first of its kind to comprehensively apply the Whisper model to emotion classification in cross-database scenarios.

8.2 Proposed Work

8.2.1 Web-scale Supervised Pertraining for Speech Recognition (WSPSR)

Whisper is a novel ASR model which was introduced recently in September 2022. Its name, WSPSR, stands for "Web-scale Supervised Pretraining for Speech Recognition," a multitask and multilingual model. Unlike previous models, such as Wav2Vec that rely on pre-training on unlabelled audio data [36], Whisper has been trained on a massive dataset of labeled audio data collected from the Web. The architecture of Whisper emphasizes the significance of using a diverse and extensive supervised dataset for training, which can significantly improve the system's performance and adaptability to new data. The architecture design of Whisper is illustrated in Figure 8.2 as described in [60].

Table 8.1: Details of Different Whisper Models. After [65].

Model	**Layers**	**Dimension of Output Vector**	**Parameters**
Tiny	4	$1 \times 1500 \times 384$	39 M
Base	6	$1 \times 1500 \times 512$	74 M
Small	12	$1 \times 1500 \times 768$	244 M
Medium	24	$1 \times 1500 \times 1024$	769 M
Large	32	$1 \times 1500 \times 1284$	1550 M

As shown in Figure 8.3, the current study suggests a transfer learning-based strategy for emotion categorization using deep neural network (DNN) classifiers, such as CNN and ResNet. Depending on the Whisper model used, the Whisper Encoder Module requires a Mel spectrogram of the audio signal, which is padded to 30 seconds. It then produces a fixed-dimensional vector. Instead of sending this fixed-dimensional vector to the Whisper Decoder Module, different DNNs use it as a feature for classification. Figure 8.1 shows the log-Mel-Spectrogram utilized in this study to represent various emotions.

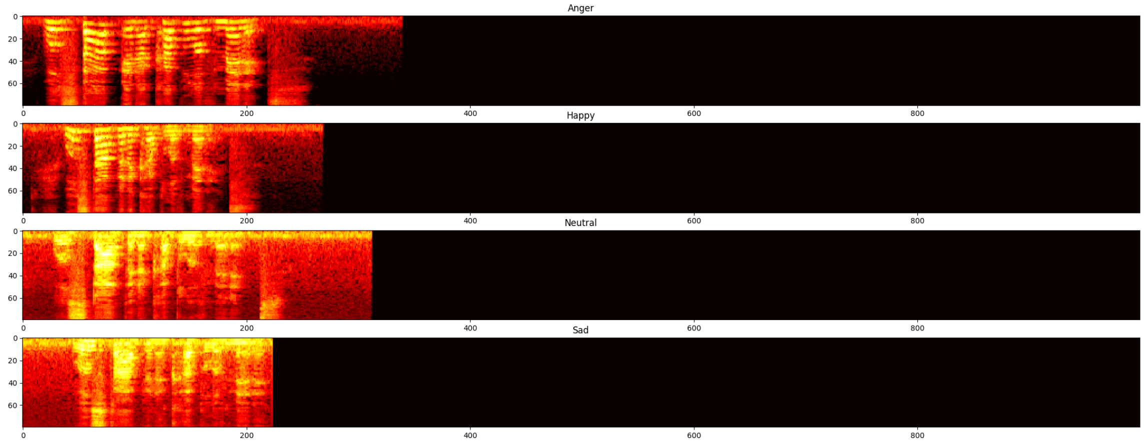


Figure 8.1: Log Mel-Spectrogram of the same sentence in different emotions after padded by Whisper model (to make all utterances of 30 seconds length).

The technical details of numerous Whisper Model variations, each with a different number of trainable parameters [65], are listed in Table 1. However, the experiments in this paper were limited to the *base* model because of resource limitations.

8.3 Experimental Details

As mentioned in Chapter 3, three datasets, namely, EmoDB, ESD, and CREMA-D are used in this study. To ensure reliable results, this study used different sets of speakers for training, validation, and testing in both the datasets. We also maintained a similar ratio of train-test-validation, close to 70 %-10 %-20 % for both ESD and CREMA-D (70-10-20 in ESD and 70.51-9.73-19.75 in CREMA-D). In addition, the EmoDB dataset was exclusively utilized for testing purposes in this study, as it contains a smaller number of files (339 files for four emotions). For this work, we have used *13*-D static features for both MFCC and GFCC with 16 *kHz* sampling rate, and a window length of 25 ms. The classifier details are already mentioned in Chapter 3.

8.4 Experimental Results

The main aim of the preliminary experiments was to determine a neural network architecture that could effectively classify the extracted features into different emotions with high accuracy. Upon obtaining promising results on the ESD dataset, the same architecture, as depicted in Figure 8.3, was utilized for the

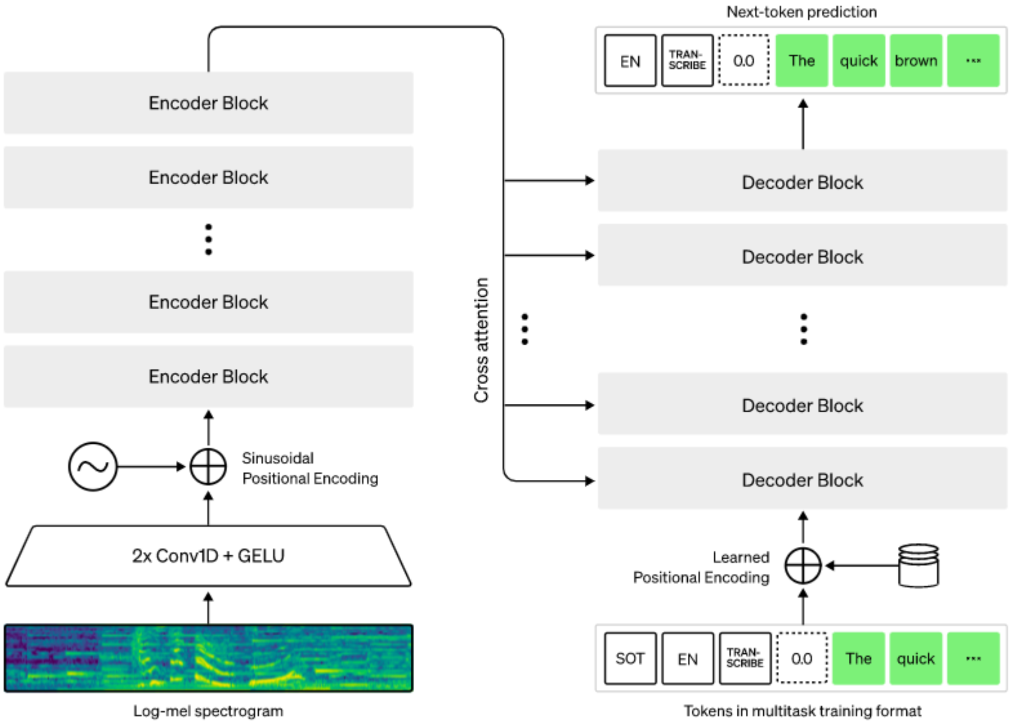


Figure 8.2: Complete Whisper Model. After [60].

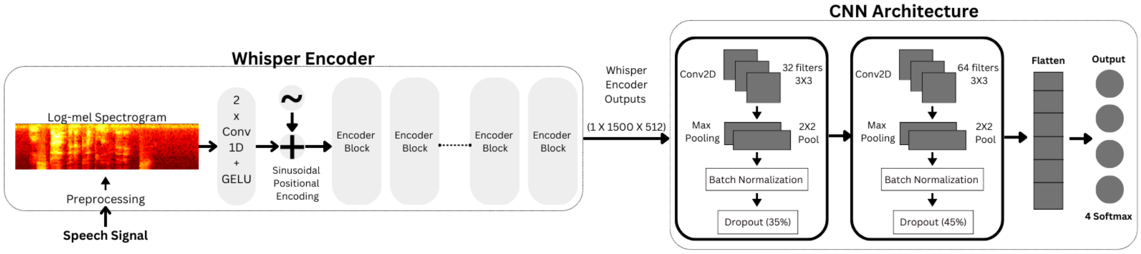


Figure 8.3: Pipeline of architecture used for SER task.

CREMA dataset. The findings revealed that the identical architecture worked effectively with Whisper features on both the datasets. The outcomes, comprising accuracy and other statistical metrics, are presented in Table 8.2 and Table 8.3. It is noted that whisper features outperform baseline features MFCC and GFCC by **20** % on CNN and **8-30** % using ResNet. As seen in the literature, prosodic features extensively effective for emotion classification are generally *suprasegmental* in nature, and the baseline features capture *segmental* information. Since whisper features capture around 30 s of utterance, and hence, it is able to capture the prosodic elements much better than state-of-the-art MFCC and GFCC features. Emotion is not dependent on a single speech segment, however, it is understood by the entire utterance; using long-term features helps capture better. Also, due to

Table 8.2: Results for different features using CNN classifier for matched and mismatched conditions. After [88].

Train Dataset	Test Dataset	Feature Set	Accuracy (in %)	F1 Score	MCC	Jaccard Index	Hamming Loss
ESD	**ESD**	**MFCC**	49.21	0.4701	0.3417	0.3166	0.5079
		GFCC	50.86	0.4630	0.3698	0.3254	0.4914
		Whisper	**86.86**	**0.8595**	**0.8333**	**0.7630**	**0.1314**
	Crema	**MFCC**	30.4	0.2051	0.0958	0.1295	0.696
		GFCC	45.91	0.4374	0.3218	0.2967	0.5409
		Whisper	**56.18**	**0.4945**	**0.4493**	**0.3524**	**0.4382**
	EmoDB	**MFCC**	62.83	0.6221	0.4979	0.4676	0.3709
		GFCC	37.46	0.2046	0.0086	0.1408	0.6254
		Whisper	**56.93**	**0.5003**	**0.4606**	**0.3697**	**0.4307**
Crema	**ESD**	**MFCC**	24.86	0.1417	-0.0027	0.0867	0.7514
		GFCC	31.71	0.2681	0.0990	0.1696	0.6829
		Whisper	**51.78**	**0.4482**	**0.3917**	**0.3103**	**0.4821**
	Crema	**MFCC**	51.36	0.5118	0.3562	0.3497	0.4864
		GFCC	55.97	0.5468	0.4457	0.3850	0.4403
		Whisper	**72.53**	**0.7243**	**0.6477**	**0.5680**	**0.2746**
	EmoDB	**MFCC**	51.33	0.4611	0.3218	0.3263	0.4867
		GFCC	40.71	0.2665	0.1348	0.1815	0.5929
		Whisper	**60.18**	**0.5987**	**0.4521**	**0.4400**	**0.3982**

huge multi-lingual training data, whisper features effectively capture emotional context without the hindrance of cultural and linguistic barriers.

8.5 Chapter Summary

This study investigated features obtained from the Whisper Encoder for SER task. This model has demonstrated accuracy on par with that of humans in ASR and is trained on a large amount of audio data. We propose that the superior performance of the Whisper features is due to their ability to capture both *acoustic* and *linguistic* information, encompassing all relevant emotional aspects of the audio speech signal. Results from experiments conducted showed that the Whisper model performed exceptionally well in both classifiers and achieved significantly higher accuracy than traditional feature sets on both datasets. We further validated their results by employing additional statistical parameters and analyzing the confusion matrix. All experimental conditions consistently showed that the proposed Whisper Encoder-based feature set outperformed other feature sets in all metrics evaluated, providing strong evidence supporting the proposed approach. In the next chapter, we discuss an exploratory work where the LFRCC feature on infant cry classification is studied.

Table 8.3: Results for different features using ResNet classifier for matched and mismatched conditions. After [88].

Train Dataset	**Test Dataset**	**Feature Set**	**Accuracy (in %)**	**F1 Score**	**MCC**	**Jaccard Index**	**Hamming Loss**
ESD	**ESD**	**MFCC**	45.07	0.3767	0.2924	0.2584	0.5493
		GFCC	49.14	0.4596	0.3399	0.3111	0.5086
		Whisper	**79.85**	**0.7863**	**0.7464**	**0.6589**	**0.2014**
	Crema	**MFCC**	33.12	0.2633	0.1374	0.1647	0.6688
		GFCC	45.70	0.4383	0.3041	0.2975	0.5430
		Whisper	**46.54**	**0.4039**	**0.3178**	**0.2668**	**0.5346**
	EmoDB	**MFCC**	51.62	0.2198	0.0391	0.1481	0.6224
		GFCC	37.76	0.3800	0.3611	0.2811	0.4838
		Whisper	**55.75**	**0.4894**	**0.4281**	**0.3520**	**0.4424**
Crema	**ESD**	**MFCC**	23.43	0.2037	-0.0227	0.1191	0.7657
		GFCC	41.57	0.4007	0.2250	0.2756	0.5843
		Whisper	**49.14**	**0.3675**	**0.3861**	**0.2639**	**0.5086**
	Crema	**MFCC**	50.10	0.4389	0.3632	0.3066	0.4989
		GFCC	55.97	0.5431	0.4197	0.3862	0.4403
		Whisper	**70.65**	**0.7243**	**0.6477**	**0.568**	**0.2746**
	EmoDB	**MFCC**	40.12	0.2886	0.1214	0.1800	0.5988
		GFCC	37.46	0.2097	0.0136	0.1430	0.6254
		Whisper	**48.90**	**0.3776**	**0.2468**	**0.2568**	**0.5309**

CHAPTER 9
Exploratory Work: LFRCC for Infant Cry Classification

9.1 Introduction

Infant cry research is interdisciplinary in nature involving paediatrics, cognition, psychology, engineering, language acquisition, robotics, prosody, and autism spectrum disorders (ASD) [29]. For example, language acquisition research has shown that infants have remarkable ability to distinguish two different languages (including native *vs.* non-native) within just four days of birth and thus, deeper research in language acquisition may help for developing better speech recognition and speech understanding systems, such as in robotics. In addition, cry units of infants who later diagnosed with ASD are also found to have higher fundamental frequency (F_0) than controls similar to pathological infant cry, thereby making it challenging to identify acoustic cues of normal, ASD, and pathological cry samples.

Around 3 million infants die within first four months of birth due to various reasons, such as pathology, malnutrition, vaccine preventable disease, abnormalities in the brain stem controlling breathing function, etc. In this context, infant biometrics using fingerprint and cry signal are developed [29]. In the context of pathologies, birth asphyxia and related abnormalities, in particular, sudden infant death syndrome (SIDS) are leading cause of death for infants [53]. Landmark investigations sponsored by the National Institute of Health (NIH), USA, reported evidences of abnormalities in brainstem (in particular, medulla oblongata) that is known to control breathing functions, for the infants who died of SIDS [11]. Further, clinical diagnosis of asphyxia is logistics heavy and costly and thus, it is mostly diagnosed late, however, by then, severe neurological damage would have already occurred to the infants [28]. Further, acoustic cues of the deaf infant cry are depends upon hearing loss, type and duration of rehabilitation, and the age of pathology detection [59]. Moreover, not every infant has a luxury of being taken care by a Neonatal Intensive Care Unit (NICU) and a team of paediatricians, more so, in the Indian context. To that effect, several attempts to develop assistive technologies, such as baby cry analyzer [2], baby pod [3], and Ubenwa mobile app indicates a genuine need to develop a cost-effective and non-invasive cry diagnosis tool as a supplement to well known *Apgar* count (that is a function of baby weight, preterm *vs.* fullterm, cry being vigorous *vs.* shill, etc.). In this context, this work investigates signal processing-based approach for infant cry classification, where asphyxia and deaf are considered as pathological cry signals.

Even though research on infant cry analysis started as early as 1960's and that this problem is socially relevant, the progress in this field is slower primarily due to several challenges. In particular, ethical issues associated with data collection, higher fundamental or pitch frequency (F_0) and hence, infant cry signal suffers from poor spectral resolution. Original investigations in [92] identified ten distinct cry modes to indicate differences in the manner of modes of vibrations associated with the vocal folds voicing, i.e., F_0 and its harmonics ($kF_0, k \in Z$) via narrowband spectrogram (having window duration less than a pitch period, i.e., $\sim$ 1-2 ms [64]). However, these investigations were limited to only normal infant cries and later extended to analysis of asthma, Hypoxic Ischemic Encephalopathy (HIE), and larynx abnormalities . These studies exploited narrowband spectrogram due to their capability of reflecting manner of variations in F_0 and kF_0, where formant structures are difficult to observe due to quasi-periodic sampling of vocal tract spectrum by high pitch-source harmonics (under the assumption of linear time-invariant cry production model). Recently, MFCC features modeled using statistical classifier, namely, GMM are used for this task [8]. However, this work investigates the excitation source-based features based on the classic Linear Prediction (LP) concept for infant cry analysis. Since cries are very difficult to distinguish for the human ear, excitation source-based features, which capture the characteristics of glottal airflow needed for sound (cry) production, proved effective. This study also investigated the effect of LFRCC on cross-database (i.e., *mismatched* conditions) and combined database evaluation scenarios. It was observed that LFRCC outperformed MFCC and LFCC by 24.9% and 17.43%, respectively, for mismatched conditions and by $0.27\% - 1.11\%$ for the combined database. The block diagram of the proposed feature is given in Figure 9.1.

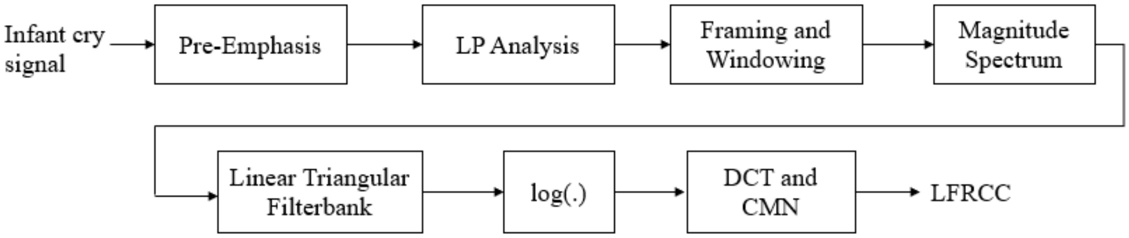


Figure 9.1: Schematic block diagram of LFRCC feature extraction. After [81].

9.2 Database Details

Baby Chillanto database used for this work was originally developed by the recordings conducted by medical doctors, which is a property of NIAOE-CONACYT, Mexico [68], [69]. Table 9.1 shows the statistics of the Baby Chilanto database with a total of 1049 healthy cries and 1219 pathological cries. Another database used was the In-house database. It was collected by [20],[24]. Table 9.1 shows the statistics of the In-house database with a total of 793 healthy cries and 416 pathological cries. For a fair comparison, each cry segment was resampled at a uniform sampling frequency of 16 *kHz*.

Table 9.1: Number of Cry Recordings in Baby Chilanto and In-house Corpora. After [20], [24], [68], and [69].

Class	**Category**	**Baby Chilanto**	**In-House**
Healthy	Normal	507	793
	Hunger	350	-
	Pain	192	-
Pathology	Asphyxia	340	215
	Deaf	879	-
	Asthma	-	182

9.3 Baseline Used

The performance of the proposed LFRCC feature set was compared with state-of-the-art feature sets, namely, MFCC and LFCC. 39-*D* MFCC and 39-*D* LFCC features were extracted using a window length of 30 ms and window overlap of 15 ms. Each containing 13-*D* static + 13-Δ + 13-ΔΔ features. The GMM and CNN classifiers' architecture details are discussed in Chapter 3.

9.4 Experimental Results

9.4.1 Effect of LP Order

In this sub-Section, the results obtained by varying LP order is analyzed. Figure.9.2 shows the effect of LP order on both datasets. It can be observed from Figure.9.2 that % classification accuracy decreases as we increase the LP order. The optimal results are obtained keeping LP order 4, where we get 99.03% and 91.50 % accuracy on Baby Chilanto and In-House datasets, respectively. Source modelling is not working for higher LP orders as the vocal tract of babies is very small as compared to adults and has less mass of the vocal folds. For infants, the glottal cycle (i.e., closed phase, open phase, and return phase) is of relatively lesser duration ($\sim$ 1-2 ms) as compared to adults and the contextual information is not needed for infant cry analysis and the higher order predictor memory might confuse the model. Further, due to the much lesser length of the vocal tract (< 5cm), lesser predictor memory is required to model it [13].

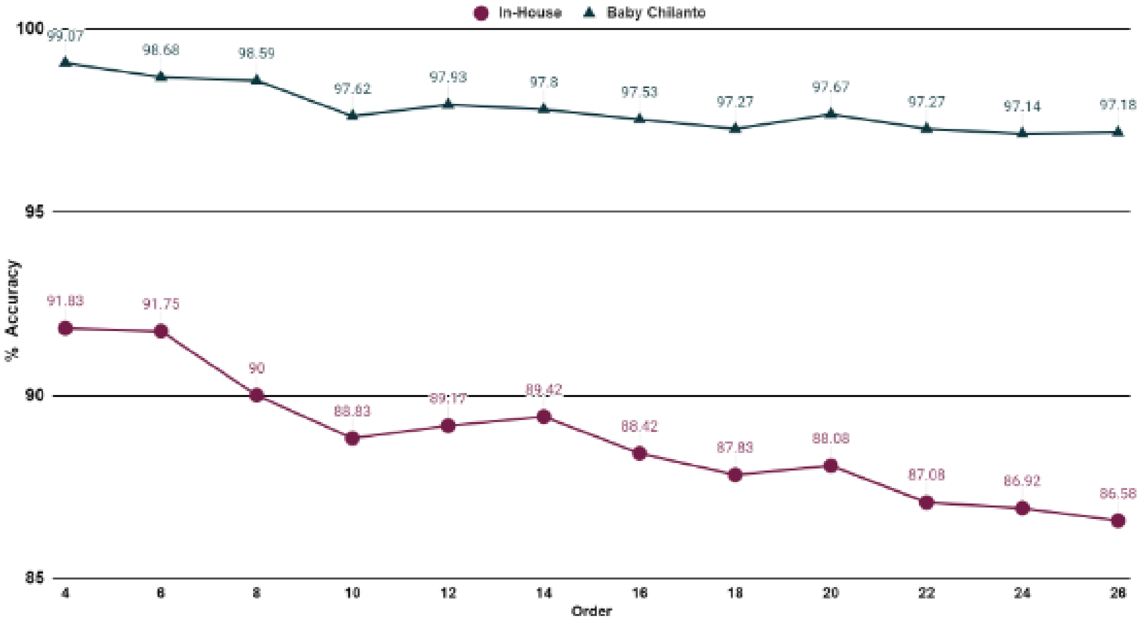


Figure 9.2: Effect of LP order for LFRCC using GMM classifier. After [84].

9.4.2 Results for Matched Conditions

Table 9.2 and Table 9.3 shows the classification accuracy of proposed LFRCC features on GMM and CNN classifiers, respectively. From Table 9.2, it can be inferred that LFRCC gives comparable results with baseline MFCC and LFCC (99.21 %

Table 9.2: Classification Accuracy (%) of MFCC, LFCC, and LFRCC using GMM Classifier. After [84].

Database	MFCC	LFCC	LFRCC
Baby Chillanto	99.47	99.25	99.21
In-house Corpus	95.75	94.75	91.83
Combined	96.43	96.96	**97.23**

Table 9.3: Classification Accuracy (%) of MFCC, LFCC, and LFRCC using CNN Classifier. After [84].

Database	MFCC	LFCC	LFRCC
Baby Chilanto	98.59	97.88	97.02
In-house Corpus	87.91	84.56	**88.68**
Combined	97.12	97.56	**98.23**

and 91.83 % for Baby Chilato and In-House corpus, respectively). LFRCC outperforms MFCC and LFCC by 0.78% and 0.27%, respectively. This is because LFRCC, being excitation source-based features, it is able to identify the difference in characteristics of glottal airflow used for cry production for different pathologies. The increase in number of training and testing samples also help in classifying cries better.

Table 9.3 shows the results from deep neural network-based CNN classifier, where the proposed feature outperforms MFCC and LFCC by 0.77% and 4.12% for in-house corpus. Infant cries have high-pitched cries, which results in a greater role of factor [64]. This results in rich discriminative acoustic cues in the lower frequency regions. MFCC has a better resolution in lower-frequency regions than LFCC. This results in better performance of MFCC than LFCC. These results indicate that the excitation source-based features help in the binary classification of infant cry because the glottal airflow needed for different cry categories is different, thus, indicating its potential for this study.

9.4.3 Results for Mismatched Conditions

Table 9.4 highlights the classification accuracy for cross-database (CD) or mismatched conditions. Here, CD1 implies training with Baby Chilanto database and testing on In-House corpus. In CD2, training is done on In-House corpus and testing on Baby Chilanto database. It is observed that the proposed LFRCC features work well for cross-database classification. For CD1, LFRCC gave the best performance as all the categories of cries on testing were present in training (normal, asphyxia, and asthma), which helps in classification as the proposed features captured the breathing and in turn, the sound production (cry) better for all the classes. The classification accuracy drops to 56.34 % in LFRCC because, the proposed feature is capable of extracting the source-based information and the categories of cries used in testing (hunger, pain, deaf) are different than that in training. MFCC does not perform well as it is based on human sound perception and the cries used have minimal difference when heard and thus, making it difficult to differentiate.

Table 9.4: Classification Accuracy (%) of MFCC, LFCC, and LFRCC using GMM Classifier. After [84].

Database	**MFCC**	**LFCC**	**LFRCC**
CD1	31.08	39.67	**60.50**
CD2	35.95	59.74	56.34

9.5 Chapter Summary

In this study, an excitation source-based feature, namely, LFRCC was used for infant cry classification. Vocal tract-based cepstral features MFCC and LFCC were used for performance comparison. The objective was to use the complementary information from source-based features for infant cry classification. The results showed comparable performance with traditional state-of-the-art (MFCC and LFCC) features. Though, it was found that LFRCC is useful for cross-database classification due to its focus on the source (glottal airflow) needed for sound (cry) production. Its ability to classify cries that are, in general, very hard to distinguish for humans and auditory-based spectral features is the motivation for this study. In the next chapter, we focus on developing a local API for SER task.

CHAPTER 10

Local API for SER

10.1 Introduction

The motive of this thesis is to aid in SER problem. To understand the efficiency of proposed CQPC feature, an Application Programming Interface (API) was developed. An API for SER promotes collaboration and innovation in the field. By making the technology easily accessible to a wider community of developers, it encourages the exchange of ideas, the development of new applications, and the advancement of SER techniques. This collective effort can lead to significant improvements in speech emotion analysis, benefiting various industries, such as healthcare, customer service, and entertainment.

Furthermore, an API for SER facilitates the creation of personalized user experiences. Applications can leverage the emotional insights extracted from speech to adapt their responses or tailor their content accordingly. For instance, a virtual assistant can adjust its tone or provide empathetic responses based on the detected emotions, resulting in a more engaging and human-like interaction.

Moreover, building an API allows for scalability and extensibility. As SER technology evolves, developers can easily update their applications by integrating newer versions of the API, ensuring access to the latest advancements in emotion recognition. This flexibility enables developers to stay at the forefront of SER research and deliver enhanced user experiences to their customers.

Under the NLTM: BHASHINI project for Assistive Speech Technologies (AST), we built as API for Emotion Classification, Infant Cry Classification, and Dysarthric Speech Classification. The details of the same are discussed below.

10.2 Model for SER

To develop an API, it is important to understand the front end and back end functionality. The front end of an API refers to the part that users interact with directly.

It involves the user interface components, such as forms, buttons, and visual elements, which enable users to send requests to the API. This can be implemented in various software, including web pages and mobile applications, that utilize the API for requesting and receiving data.

In contrast, the back end of an API handles the server-side processing and logic. It receives incoming requests from the front end, processes them, and generates appropriate responses. The back end performs operations like data retrieval, manipulation, validation, and additional business logic required to fulfill the requested actions. It also deals with tasks, such as connecting to databases, performing calculations, and ensuring authentication and security measures.

The front end and back end working of our API is explained below.

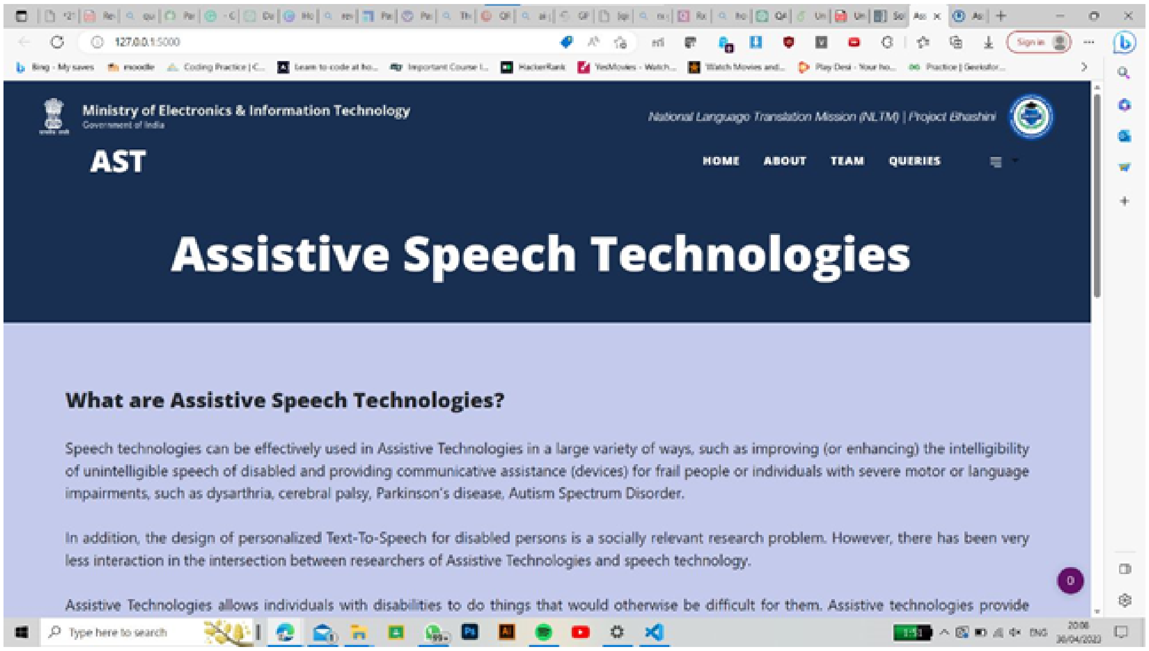
10.2.1 Front end

CSS, JavaScript, and HTML are fundamental tools used in front-end development to create captivating and interactive user interfaces. They were used to develop the model shown below. HTML, short for HyperText Markup Language, serves as the structural backbone of a web page. It defines the various elements and content present on the page, including headings, paragraphs, images, and links. By utilizing HTML tags, developers establish the hierarchical structure of the content, allowing web browsers to render it correctly. HTML provides the foundation and framework upon which CSS and JavaScript can operate by defining the layout and fundamental structure of a web page.

CSS, or Cascading Style Sheets, complements HTML by managing the presentation and styling of the elements defined in the HTML markup. Through CSS, developers have the ability to specify the visual appearance, layout, and visual effects of these elements. It empowers developers to define properties such as colors, fonts, sizes, margins, and positioning, enabling precise control over the overall look and feel of the user interface. CSS ensures consistency, responsiveness, and customization, facilitating the creation of visually appealing and user-friendly designs.

JavaScript, a versatile programming language, brings interactivity and dynamic functionality to the front end of a website or application. With JavaScript, developers can incorporate interactive behaviors, animations, and real-time updates into the user interface. It allows for the handling of user interactions, validation of input, manipulation of the Document Object Model (DOM), retrieval of data through AJAX requests, and the creation of intricate interactive features. JavaScript plays a vital role in enhancing user experiences and extends the capabilities of HTML and CSS by providing advanced functionality beyond their individual scope.

- STEP I: The website for Assistive Speech Technologies is shown below. The motivation and need for this API are briefed on this page, along with the host organization (DA-IICT) and funding body, Ministry of Electronics and Information Technology (Meity) that sponsored this consortium project (Figure 10.1).



127.0.0.1:5000
Ministry of Electronics & Information Technology
Government of India
National Language Translation Mission (NLTM) | Project Bhashini
AST
HOME
ABOUT
TEAM
QUERIES
Type here to search

Assistive Speech Technologies

What are Assistive Speech Technologies?

Speech technologies can be effectively used in Assistive Technologies in a large variety of ways, such as improving (or enhancing) the intelligibility of unintelligible speech of disabled and providing communicative assistance (devices) for frail people or individuals with severe motor or language impairments, such as dysarthria, cerebral palsy, Parkinson's disease, Autism Spectrum Disorder.

In addition, the design of personalized Text-To-Speech for disabled persons is a socially relevant research problem. However, there has been very less interaction in the intersection between researchers of Assistive Technologies and speech technology.

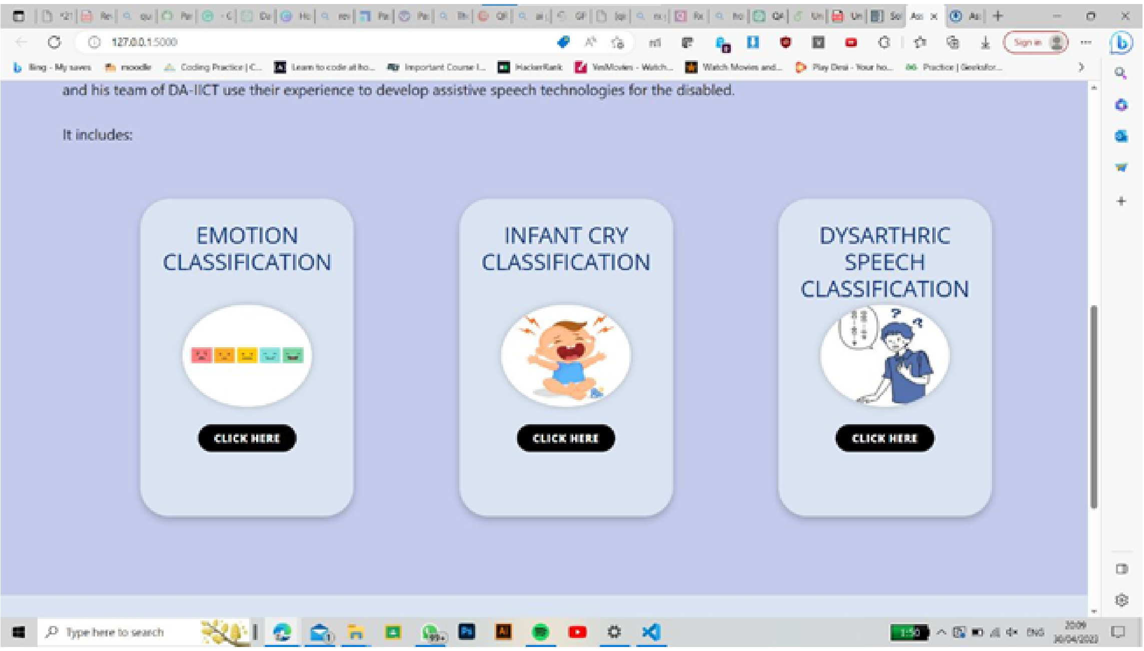
Assistive Technologies allows individuals with disabilities to do things that would otherwise be difficult for them. Assistive technologies provide

Figure 10.1: Assistive Speech technology website.

- STEP II: Scrolling the page gives three icons, namely: Emotion Classification, Infant Cry Classification, and Dysarthric Speech Classification. Press the button "**CLICK HERE**" to view the desired page (see Figure 10.2).

- STEP III: The motivation for developing Emotion Classification is described on this page. The CNN model is employed to classify five emotions, namely, anger, happiness, neutral, sadness, and surprise. This model is trained by extracting Constant Q Pitch Coefficients (CQPC) features from the audio file (refer Figure 10.3).

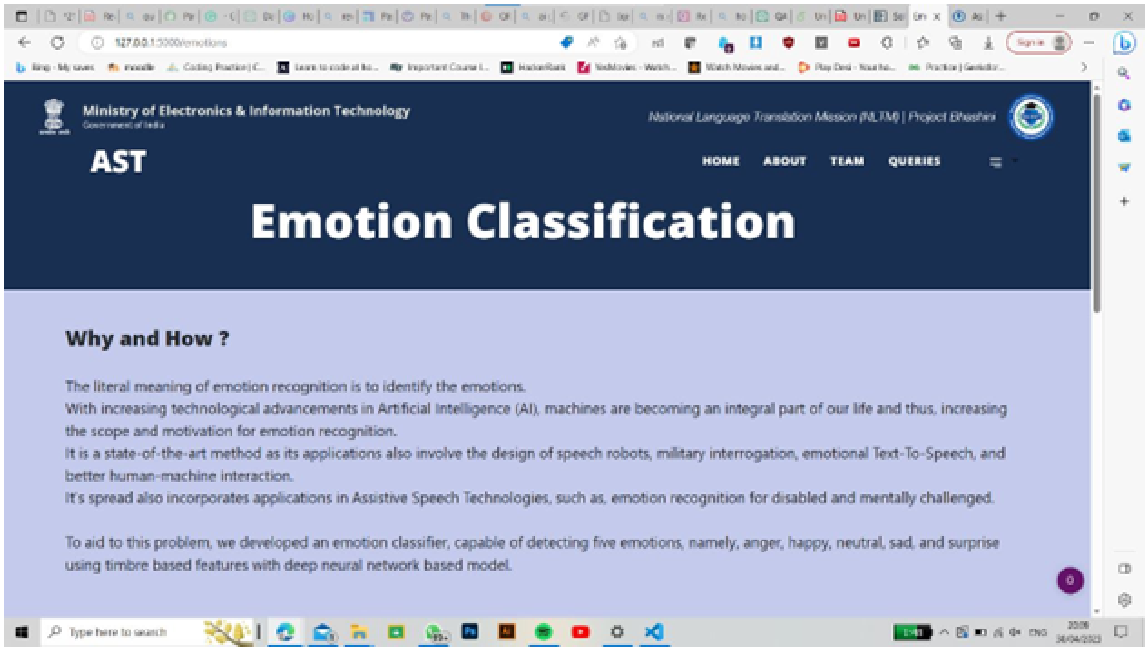
- STEP IV: Scrolling the page, the icon "**Choose File**" is seen. Press this to upload the desired audio file to test/classify. After choosing the desired file, click "**Classify**" to get the emotion in the chosen audio. The emotion along with its emoji is displayed (see Figure 10.4).



and his team of DA-IICT use their experience to develop assistive speech technologies for the disabled.

It includes:

Figure 10.2: API front end for Assitive Speech Technology.


127.0.0.1:5000/emotions
Ministry of Electronics & Information Technology
Government of India
National Language Translation Mission (NLTM) | Project Bhashini
AST
HOME
ABOUT
TEAM
QUERIES

Emotion Classification

Why and How ?

The literal meaning of emotion recognition is to identify the emotions.
With increasing technological advancements in Artificial Intelligence (AI), machines are becoming an integral part of our life and thus, increasing the scope and motivation for emotion recognition.
It is a state-of-the-art method as its applications also involve the design of speech robots, military interrogation, emotional Text-To-Speech, and better human-machine interaction.
It's spread also incorporates applications in Assistive Speech Technologies, such as, emotion recognition for disabled and mentally challenged.

To aid to this problem, we developed an emotion classifier, capable of detecting five emotions, namely, anger, happy, neutral, sad, and surprise using timbre based features with deep neural network based model.

Figure 10.3: Emotion Classification Page.

10.2.2 Back end

Flask, a widely-used back end framework in web development that works seamlessly with front-end technologies, such as HTML, CSS, and JavaScript is used for this work. This lightweight and adaptable framework, written in Python, is specifically designed to facilitate the rapid and efficient construction of web ap-

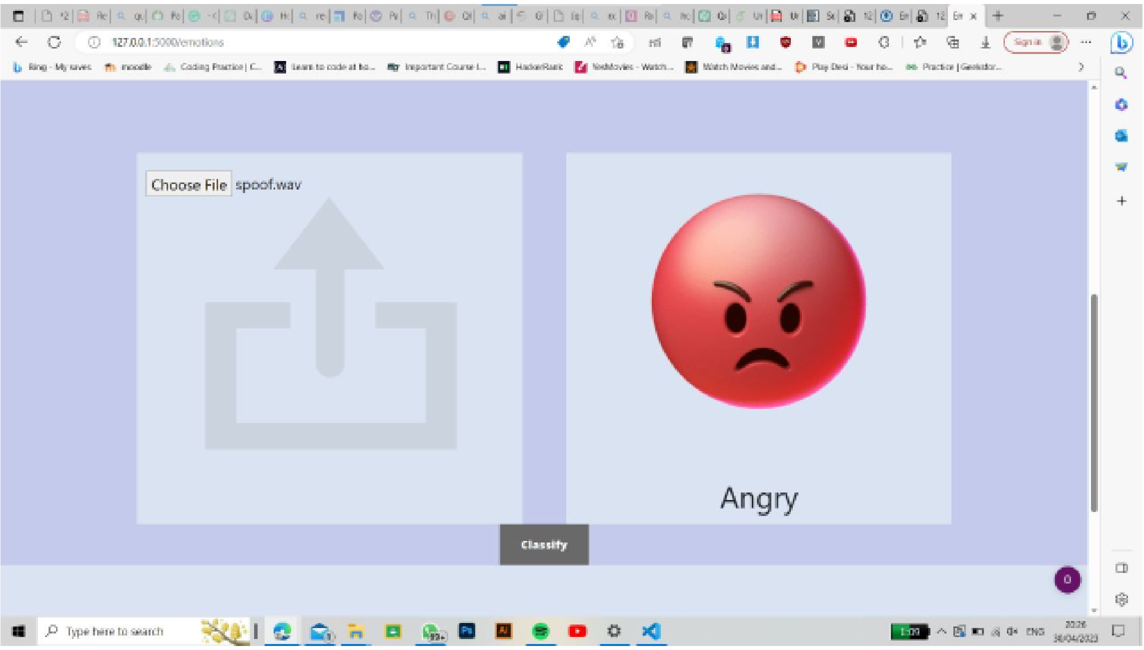


Figure 10.4: Output image.

plications.

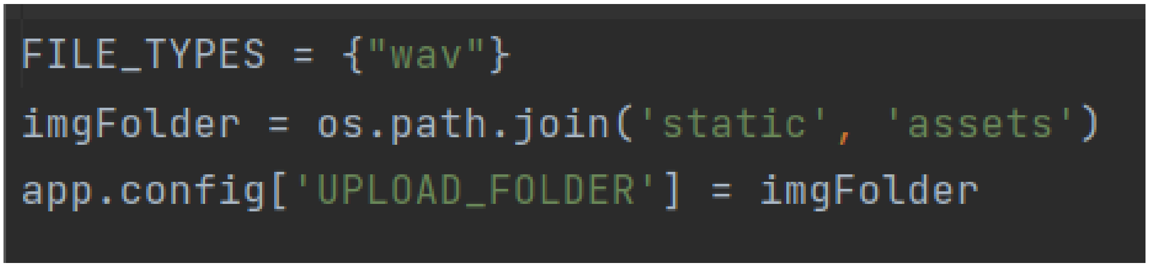
We call the models by clicking on the submit button after entering the *.wav* files for each speech assistive technology and then return the result we obtain and output it on the front end (screen).

Calling the model function and linking the corresponding images:

In the initial step, the characteristics and compatibility of given *.wav* files will be examined. If the necessary requirements are met, a model function will be invoked with the user-provided input. This function will load the model from the same folder/directory as the *main.py* file (ensure both the files are kept together i.e., having the same path or directory). The function will generate a numerical output, where each number corresponds to a specific image or GIF. Subsequently, the output will be displayed on the screen.

- Step I: We have established two folders (static and templates) alongside a *main.py* file that encompasses the back end code. The Static folder houses essential components, such as CSS and JavaScript scripts, as well as the necessary assets including images or logos that are utilized in the project. The templates folder contains HTML files for each page.

- Step II: To begin, we will initiate the process by locating the path to the assets folder, where the desired final images are stored.

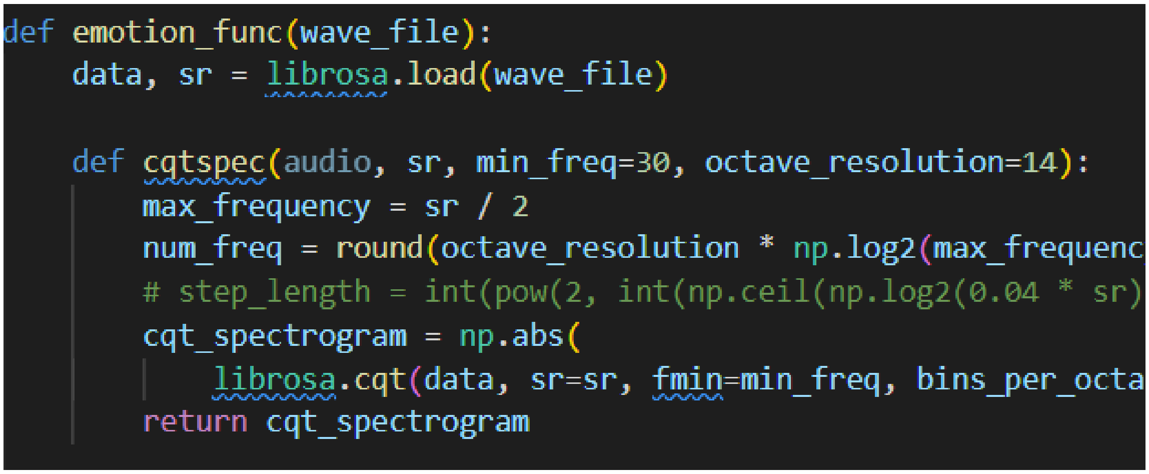


```
FILE_TYPES = {"wav"}
imgFolder = os.path.join('static', 'assets')
app.config['UPLOAD_FOLDER'] = imgFolder
```

Figure 10.5: Code to access folders.

In Figure A.5 FILE_TYPES contains the valid input types' names, "os.path.join" is used to set the path to assets folder. The address is then passed and saved to application configuration folder ('UPLOAD FOLDER').

- Step III: Following that, proceed to define the AST function specific to this context, which is referred to as "emotion_func." Within this function, the input file obtained from the front end, identified as ".wav file" in our case, is passed as the input parameter (refer Figure 10.6).

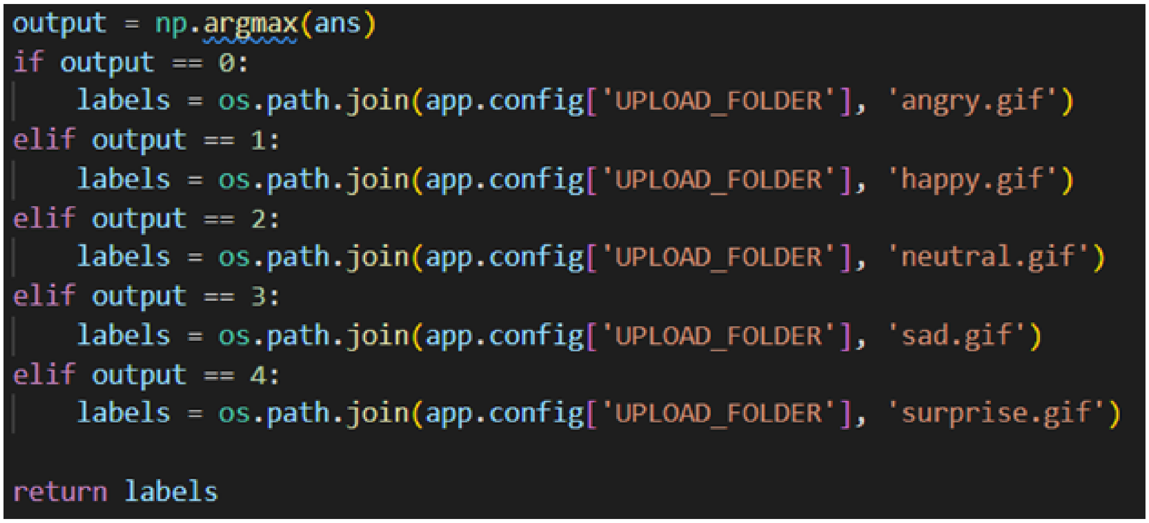


```
def emotion_func(wave_file):
    data, sr = librosa.load(wave_file)

    def cqtspec(audio, sr, min_freq=30, octave_resolution=14):
        max_frequency = sr / 2
        num_freq = round(octave_resolution * np.log2(max_frequenc
        # step_length = int(pow(2, int(np.ceil(np.log2(0.04 * sr)
        cqt_spectrogram = np.abs(
            librosa.cqt(data, sr=sr, fmin=min_freq, bins_per_octa
        return cqt_spectrogram
```

Figure 10.6: Emotion function.

- Step IV: Afterward, there is an association between emotions and numerical values, requiring us to retrieve corresponding images. These images are stored in the "static/assets/<image name>" directory. To access them, we have already stored the path up to the assets folder in the app.config ("UPLOAD_FOLDER"). By appending the file name to this path, we can access the specific file associated with the corresponding emotion. Utilizing the os.path.join method, we combine and create a unified path, storing them in the "labels" variable. Finally, we return the labels containing the paths to the respective images (Figure 10.7).

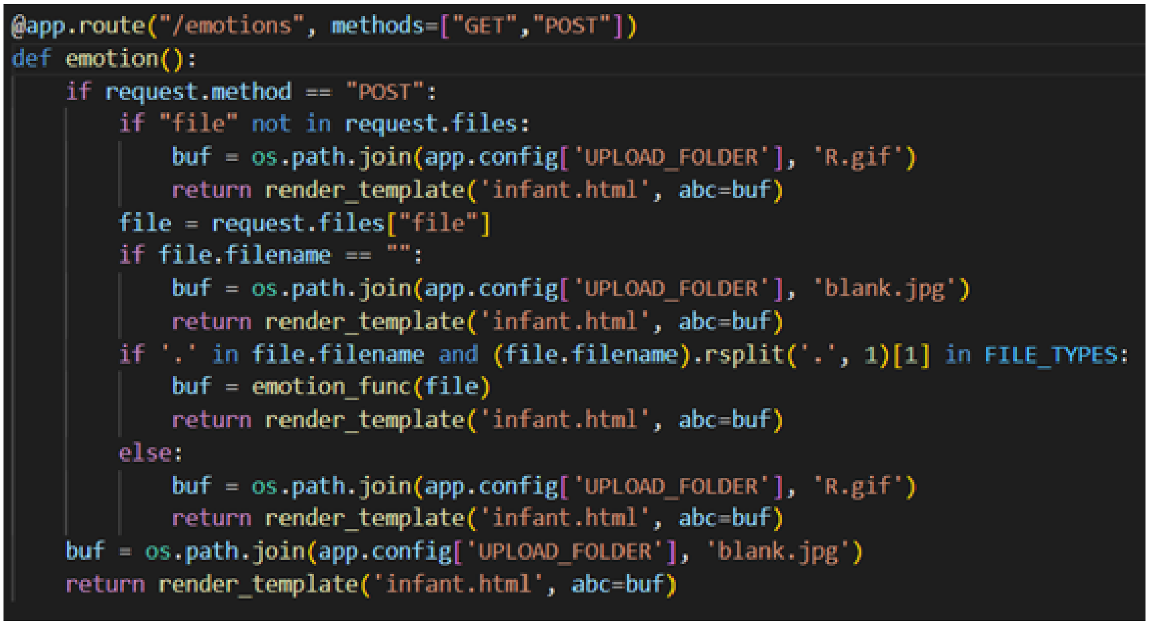


```
output = np.argmax(ans)
if output == 0:
    labels = os.path.join(app.config['UPLOAD_FOLDER'], 'angry.gif')
elif output == 1:
    labels = os.path.join(app.config['UPLOAD_FOLDER'], 'happy.gif')
elif output == 2:
    labels = os.path.join(app.config['UPLOAD_FOLDER'], 'neutral.gif')
elif output == 3:
    labels = os.path.join(app.config['UPLOAD_FOLDER'], 'sad.gif')
elif output == 4:
    labels = os.path.join(app.config['UPLOAD_FOLDER'], 'surprise.gif')

return labels
```

Figure 10.7: Code for Step IV.

- Step V: To facilitate navigation or create new pages, we utilize the "@app.route (the name of the desired page)" decorator. If there is a need to handle user input and provide output or store the input, we specify the methods=["GET", "POST"]. When a submit button is clicked, it triggers a POST request. Leveraging this information, we have developed the "emotion" function. Within this function, we check if a POST request has been made, indicating that a submit button was clicked. In this case, we verify whether any files were uploaded by the user. If no files were uploaded, we return a blank.jpg or R.gif image.

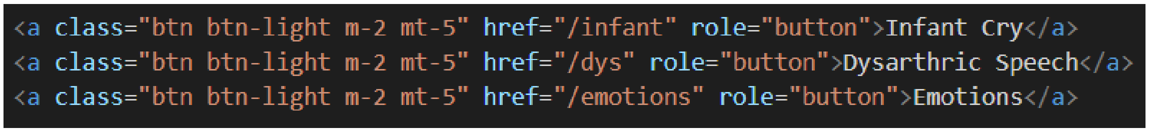


```
@app.route("/emotions", methods=["GET","POST"])
def emotion():
    if request.method == "POST":
        if "file" not in request.files:
            buf = os.path.join(app.config['UPLOAD_FOLDER'], 'R.gif')
            return render_template('infant.html', abc=buf)
        file = request.files["file"]
        if file.filename == "":
            buf = os.path.join(app.config['UPLOAD_FOLDER'], 'blank.jpg')
            return render_template('infant.html', abc=buf)
        if '.' in file.filename and (file.filename).rsplit('.', 1)[1] in FILE_TYPES:
            buf = emotion_func(file)
            return render_template('infant.html', abc=buf)
        else:
            buf = os.path.join(app.config['UPLOAD_FOLDER'], 'R.gif')
            return render_template('infant.html', abc=buf)
    buf = os.path.join(app.config['UPLOAD_FOLDER'], 'blank.jpg')
    return render_template('infant.html', abc=buf)
```

Figure 10.8: Code for Step V.

Adding/Deleting/Changing pages:

- Step 1: First define a route of your page.
- Step 2: Then write any function you want to show in the front end or return the page using the render template module.
- Step 3: As done above, we need to mention the routes as "/emotion" or "/about" in the front end.



```
<a class="btn btn-light m-2 mt-5" href="/infant" role="button">Infant Cry</a>
<a class="btn btn-light m-2 mt-5" href="/dys" role="button">Dysarthric Speech</a>
<a class="btn btn-light m-2 mt-5" href="/emotions" role="button">Emotions</a>
```

Figure 10.9: Code for Step 3.

10.3 Chapter Summary

This chapter describes the details of building a local API for Assitive Speech Technology. The details of its building, coding, and output are shown in this chapter.

CHAPTER 11
Summary and Conclusions

11.1 Summary of the work

The objective of this thesis is to contribute to the improvement of efficient SER. Firstly, an in-depth analysis of emotions was conducted to gain insights into their characteristics. Cultural and linguistic effects on emotions were also touched upon. The literature review revealed three primary techniques for extracting features in SER: prosodic features, excitation source features, and vocal tract features. This thesis presents novel CQPC and CQHC features derived from CQT for SER as it captures the music and prosodic information well. Further, a novel LFRCC feature was introduced for capturing excitation source-based information. Phase-based features, namely MGDCC, were introduced to capture vocal tract and vocal fold state, which aided SER. Whisper features were also tested in matched and mismatched conditions for SER. Additionally, the proposed LFRCC was also tested on infant cry analysis, as both emotion and infant cry classification are highly dependent on pitch information, thus motivating to test the feature in a new problem statement. Several machine learning and deep learning models, such as GMM, CNN, ResNet, and TDNN, are employed to assess the proposed features' performance. Various evaluation metrics, including F1-score, MCC, Hamming Loss, Jaccard Index, accuracy, and confusion matrix, are used to measure the performance of these classification models. To ensure reliable and unbiased results, 5-fold cross-validation is applied, providing a robust evaluation process. These metrics serve as indicators of the accuracy and effectiveness of the models, facilitating the assessment of their suitability for real-world applications.

11.2 Conclusions

The following conclusions can be drawn from this thesis:

- Mandarin being a tonal language has higher F_0 fluctuations due to pitch variations, is louder, and has a higher energy profile than the stress-timed language English.
- Music stimulates emotions, and thus music-based features, CQPC and CQHC perform better in SER than traditional features such as MFCC and GFCC.
- Respiratory patterns' influence on emotions was studied with the help of LFRCC.
- MGDCC'S ability to capture the fine structures of spectral envelope aid in efficient SER.
- Supra-segmental information and huge multi-lingual training data aids whisper to perform well in SER.
- LFRCC's ability to focus on glottal airflow aids in cross-database infant cry classification.

11.3 Limitations of the Current Work

Some limitations of the proposed work are as follows:

- MGDCC and LFRCC features have to be tested on multiple databases to establish their performance in SER.
- Whisper experiments were limited to *base* models due to resource limitations.
- Cross-database evaluation of LFRCC on infant cry has to be tried on deep learning models.

11.4 Future Research Directions

This thesis made a humble attempt at aiding methods for better SER. Some future research directions for this problem statement are:

- Exploring techniques for generating synthetic emotional speech data to alleviate the scarcity of labeled data. Investigating data augmentation approaches such as speech style transfer, voice conversion, and emotion transfer to enhance the diversity and quality of training data.

- Considering the influence of contextual information (e.g., dialogue history, speaker characteristics, cultural background) on emotion recognition from speech. Investigating how to incorporate context into models to improve the accuracy and adaptability of SER systems.

- Investigating noise robustness qualities and latency period of LFRCC for SER.

- Using these proposed features on mentally challenged people and people suffering from Cerebral Palsy and Parkinson's disease.

- Developing novel deep learning models specifically designed for SER, considering the temporal dynamics and long-term dependencies in speech signals. Exploring advanced architectures such as RNNs, CNNs, and transformer-based models and adapting them to capture emotional cues better.

- Exploring Federated Learning for SER to maintain the privacy of people and develop a huge database for efficient SER.

List of Publications from Thesis

1. **S. Uthiraa**, Baveet Singh, Hemant Patil, “Linear Frequency Residual Features for Emotion Recognition,” **submitted** in International Conference on Pattern Recognition and Machine Intelligence (PREMI), December 12-15, Kolkata, India.

2. **S. Uthiraa**, Aditya Pusuluri, Hemant Patil, “Modified Group Delay Features for Emotion Recognition," **submitted** in International Conference on Pattern Recognition and Machine Intelligence (PREMI), December 12-15, Kolkata, India.

3. **S. Uthiraa**, Aastha Kachhi, Hemant Patil, "Linear Frequency Residual Features for Infant Cry Classification," **submitted** in International Conference on Pattern Recognition and Machine Intelligence (PREMI), December 12-15, Kolkata, India.

4. **S. Uthiraa**, Akshat Vora, Hemant Patil, "Whisper Features for Speech Emotion Recognition," **submitted** in International Conference on Pattern Recognition and Machine Intelligence (PREMI), December 12-15, Kolkata, India.

5. **S. Uthiraa**, Akshat Vora, Prathamesh Bonde, Aditya Pusuluri, Hemant A. Patil, ”Spectral and Pitch Components of CQT Spectrum for Emotion Recognition,” **rejected** in European Signal Processing Conference (EUSIPCO), Helsinki, Finland, 4-8 September 2023.

6. **S. Uthiraa**, Akshat Vora, Prathamesh Bonde, Aditya Pusuluri, Hemant A. Patil, "Combining Features from Spectral and Pitch Components of CQT Spectrum for Emotion Recognition," **rejected** in International Conference on Acoustics, Speech and Signal Processing, Greece 4-9 June 2023.

7. **S. Uthiraa**, Hemant A. Patil, "Analysis of Emotions in Speech using AESDD," **rejected** in International Conference on Speech and Computer, India, 14-16 November 2022.

8. **S. Uthiraa**, Hemant A. Patil, "Analysis of Mandarin *vs.* English Language for Emotional Voice Conversion," **rejected** in International Symposium on Chinese Spoken Language Processing, Singapore, 11-14 December 2022.

References

[1] An introduction to Tonal languages, url https://ceas.sas.upenn.edu/sites/default/files/%28talking%20drum%29introduction%20to%20tonal%20languages%20handout-william%20lodge-2007.doc#:~:text=an%20introduction%20to%20tonal%20languages&text=many%20of%20the%20languages%20of,order%20are%20in%20any%20language., Last Accessed : 2022-09-09.

[2] Baby Crying Analyzer. http://www.showeryourbaby.com/whycrbacran1.html/. {Last Accessed: 25/11/2022}.

[3] Baby Pod. https://babypod.net/en/babypod-device/. {Last Accessed: 25/11/2022}.

[4] Emotion recognition usinf CQT-based features, note = https://github.com/Badshah2507/Emotion-recognition-using-CQT-based-features/tree/main, {Last Accessed : 2022-10-27}.

[5] Multidisciplinary Media Mediated Communication, url : :http://m3c.web.auth.gr/research/aesdd-speech-emotion-recognition/, Last Accessed : 2022-08-13.

[6] open AI:Whisper, url : :https://github.com/openai/whisper, Last Accessed : 2023-05-07.

[7] Scientific American, url : :https://blogs.scientificamerican.com/observations/the-evolution-of-emotion-charles-darwins-little-known-/psychology-experiment/, Last Accessed : 2022-08-15.

[8] H. F. Alaie, L. Abou-Abbas, and C. Tadj. Cry-based infant pathology classification using gmms. *Speech Communication*, 77:28–52, 2016.

[9] M. T. Albawi, S. and S. Al-Zawi. Understanding of a convolutional neural network. *international conference on engineering and technology (ICET)*, (pp.1-6), 2017.

[10] S. B. Alex, L. Mary, and B. P. Babu. Attention and feature selection for automatic speech emotion recognition using utterance and syllable-level prosodic features. *Circuits, Systems, and Signal Processing*, 39(11):5681–5709, 2020.

[11] L. Armbrüster, W. Mende, G. Gelbrich, P. Wermke, R. Götz, and K. Wermke. Musical intervals in infants' spontaneous crying over the first 4 months of life. *Folia Phoniatrica et Logopaedica*, 73(5):401–412, 2021.

[12] A. Arvaniti. Rhythm, timing and the timing of rhythm. *Phonetica*, 66(1-2):46–63, 2009.

[13] B. S. Atal and S. L. Hanauer. Speech analysis and synthesis by linear prediction of the speech wave. *The Journal of the Acoustical Society of America (JASA)*, 50(2B):637–655, 1971.

[14] R. Bachu, S. Kopparthi, B. Adapa, and B. Barkana. Separation of voiced and unvoiced using zero crossing rate and energy of the speech signal. In *American Society for Engineering Education (ASEE) zone conference proceedings*, pages 1–7. American Society for Engineering Education, 2008.

[15] S. Basu, J. Chakraborty, A. Bag, and M. Aftabuddin. A review on emotion recognition using speech. In *2017 International conference on inventive communication and computational technologies (ICICCT)*, pages 109–114. IEEE, 2017.

[16] H. Beigi and H. Beigi. *Speaker recognition*. Springer, 2011.

[17] M. Bouchard, A.-L. Jousselme, and P.-E. Doré. A proof for the positive definiteness of the jaccard index matrix. *International Journal of Approximate Reasoning*, 54(5):615–626, 2013.

[18] I. Brata and I. Darmawan. Comparative study of pitch detection algorithm to detect traditional balinese music tones with various raw materials. In *Journal of Physics: Conference Series*, volume 1722, page 012071. IOP Publishing, 2021.

[19] J. C. Brown. Calculation of a constant Q spectral transform. *The Journal of the Acoustical Society of America, JASA*, 89(1):425–434, 1991.

[20] N. Buddha and H. A. Patil. Corpora for analysis of infant cry. *Oriental Cocosda, Vietnam*, 2007.

[21] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, B. Weiss, et al. A database of german emotional speech. In *INTERSPEECH, Lisbon, Portugal*, volume 5, pages 1517–1520, 2005.

[22] M. Cabanac. What is emotion? *Behavioural processes*, 60(2):69–83, 2002.

[23] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma. Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE Transactions on Affective Computing*, 5(4):377–390, 2014.

[24] A. Chittora and H. A. Patil. Data collection of infant cries for research and analysis. *Journal of Voice*, 31(2):252–e15, 2017.

[25] K. Dembczyński, W. Waegeman, W. Cheng, and E. Hüllermeier. Regret analysis for performance metrics in multi-label classification: the case of hamming and subset zero-one loss. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2010, Barcelona, Spain, September 20-24, 2010, Proceedings, Part I 21*, pages 280–295. Springer, 2010.

[26] M. El Ayadi, M. S. Kamel, and F. Karray. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern recognition*, 44(3):572–587, 2011.

[27] H. A. Elfenbein and N. Ambady. On the universality and cultural specificity of emotion recognition: a meta-analysis. *Psychological Bulletin*, 128(2):203, 2002.

[28] J. J. Engelsma, D. Deb, K. Cao, A. Bhatnagar, P. S. Sudhish, and A. K. Jain. Infant-id: Fingerprints for global good. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(7):3543–3559, 2021.

[29] G. Esposito and P. Venuti. Understanding early communication signals in autism: a study of the perception of infants' cry. *Journal of Intellectual Disability Research*, 54(3):216–223, 2010.

[30] T. Fawcett. An introduction to roc analysis. *Pattern recognition letters*, 27(8):861–874, 2006.

[31] K. M. Galotti. *Cognitive psychology in and out of the laboratory*. Sage Publications, 2017.

[32] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[33] R. M. Hegde, H. A. Murthy, and V. R. R. Gadde. Significance of the modified group delay feature in speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(1):190–202, 2006.

[34] I. Homma and Y. Masaoka. Breathing rhythms and emotions. *Experimental Physiology 93*, no. 9:1011–1021, 2008.

[35] M. A. Hossan, S. Memon, and M. A. Gregory. A novel approach for mfcc feature extraction. In *2010 4th International Conference on Signal Processing and Communication Systems*, pages 1–5. IEEE, 2010.

[36] Y. Iwamoto and T. Shinozaki. Unsupervised spoken term discovery using wav2vec 2.0. In *2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 1082–1086, Tokyo, Japan, 2021.

[37] R. Jerath and C. Beveridge. Respiratory rhythm, autonomic modulation, and the spectrum of emotions: the future of emotion recognition and modulation. *Frontiers in Psychology*, 11:Article 1980, August 2020.

[38] J. F. Kaiser. On a simple algorithm to calculate the'energy'of a signal. In *International conference on acoustics, speech, and signal processing (ICASSP)*, pages 381–384. IEEE, 1990.

[39] A. G. Katsiamis, E. M. Drakakis, and R. F. Lyon. Practical gammatone-like filters for auditory processing. *EURASIP Journal on Audio, Speech, and Music Processing*, 2007:1–15, 2007.

[40] R. A. Khalil, E. Jones, M. I. Babar, T. Jan, M. H. Zafar, and T. Alhussain. Speech emotion recognition using deep learning techniques: A review. *IEEE Access*, 7:117327–117345, 2019.

[41] Y. E. Kim, E. M. Schmidt, R. Migneco, B. G. Morton, P. Richardson, J. Scott, J. A. Speck, and D. Turnbull. Music emotion recognition: A state of the art review. In *Proc. ISMIR*, volume 86, pages 937–952, 2010.

[42] S. G. Koolagudi and K. S. Rao. Emotion recognition from speech: a review. *International Journal of Speech Technology*, 15(2):99–117, 2012.

[43] S. G. Koolagudi, R. Reddy, and K. S. Rao. Emotion recognition from speech signal using epoch parameters. In *2010 International Conference on Signal Processing and Communications (SPCOM)*, pages 1–5. IISc Bangalore, India, 2010.

[44] S. R. Krothapalli and S. G. Koolagudi. Characterization and recognition of emotions from speech using excitation source information. *International Journal of Speech Technology*, 16:181–201, 2013.

[45] K. J. Lang, A. H. Waibel, and G. E. Hinton. A time-delay neural network architecture for isolated word recognition. *Neural networks*, 3(1):23–43, 1990.

[46] M. D. Lewis. Bridging emotion theory and neurobiology through dynamic systems modeling. *Behavioral and brain sciences*, 28(2):169–194, 2005.

[47] T. Li and M. Ogihara. Detecting emotion in music. 2003, Johns Hopkins University.

[48] J. C. R. Licklider and I. Pollack. Effects of differentiation, integration, and infinite peak clipping upon the intelligibility of speech. *The Journal of the Acoustical Society of America*, 20(1):42–51, 1948.

[49] G. K. Liu. Evaluating gammatone frequency cepstral coefficients with neural networks for emotion recognition from speech. *arXiv preprint arXiv:1806.09010*, 2018.

[50] S. Maitlis and H. Ozcelik. Toxic decision processes: A study of emotion and organizational decision making. *Organization Science*, 15(4):375–393, 2004.

[51] J. Makhoul. Linear prediction: A tutorial review. *Proceedings of the IEEE*, 63(4):561–580, 1975.

[52] B. W. Matthews. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure*, 405(2):442–451, 1975.

[53] J. Mehler, P. Jusczyk, G. Lambertz, N. Halsted, J. Bertoncini, and C. Amiel-Tison. A precursor of language acquisition in young infants. *Cognition*, 29(2):143–178, 1988.

[54] H. A. Murthy. *Algorithms for processing fourier transform phase of signals*. PhD thesis, PhD dissertation, Indian Institute of Technology, Department of Computer . . . , 1992.

[55] H. A. Murthy and V. Gadde. The modified group delay function and its application to phoneme recognition. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).*, volume 1, pages I–68. IEEE, 2003.

[56] R. Nakatsu, J. Nicholson, and N. Tosa. Emotion recognition and its application to computer agents with spontaneous interactive capabilities. In *Proceedings of the seventh ACM International Conference on Multimedia (Part 1)*, pages 343–351, 1999.

[57] B. M. Nema and A. A. Abdul-Kareem. Preprocessing signal for speech emotion recognition. *Al-Mustansiriyah Journal of Science*, 28(3):157–165, 2018.

[58] s. ntalampiras. toward language-agnostic speech emotion recognition. *journal of the audio engineering society*, 68(1/2):7–13, january 2020.

[59] C. C. Onu, I. Udeogu, E. Ndiomu, U. Kengni, D. Precup, G. M. Sant'Anna, E. Alikor, and P. Opara. Ubenwa: Cry-based diagnosis of birth asphyxia. 31^{st} *Conference on Neural Information Processing Systems (NIPS), Long Beach, CA*, 2017.

[60] OpenAI. Introducing whisper. https://openai.com/blog/whisper/. {Last Accessed: 08/05/2023}.

[61] S. H. K. Parthasarathi, P. Rajan, and H. A. Murthy. Robustness of group delay representations for noisy speech signals. Technical report, Idiap, 2011.

[62] V. Peddinti, D. Povey, and S. Khudanpur. A time delay neural network architecture for efficient modeling of long temporal contexts. In *Sixteenth annual conference of the international speech communication association*, 2015.

[63] M. K. Pichora-Fuller and K. Dupuis. Toronto emotional speech set (TESS), 2020, Last Accessed : 2022-09-27.

[64] T. F. Quatieri. *Discrete-Time Speech Signal Processing: Principles and Practice*. 1^{st} Edition, Pearson Education India, 2015.

[65] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever. Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv:2212.04356*, 2022. {Last Accessed: 07/05/2023}.

[66] Z. Rafii. The constant-Q harmonic coefficients: A timbre feature designed for music signals [lecture notes]. *IEEE Signal Processing Magazine*, 39(3):90–96, 2022.

[67] Z. Rafii. The Constant-Q Harmonic Coefficients: A timbre feature designed for music signals [Lecture Notes]. *IEEE Signal Processing Magazine*, 39(3):90–96, 2022.

[68] O. F. Reyes-Galaviz, S. D. Cano-Ortiz, and C. A. Reyes-García. Evolutionary-neural system to classify infant cry units for pathologies identification in recently born babies. In *2008 Seventh Mexican International Conference on Artificial Intelligence, 27-31 October 2008, Atizapan De Zaragoza, Mexico*, pages 330–335. IEEE.

[69] O. F. Reyes-Galaviz, S. D. Cano-Ortiz, and C. A. Reyes-García. Validation of the cry unit as primary element for cry analysis using an evolutionary-neural approach. In *2008 Mexican International Conference on Computer Science*, pages 261–267. IEEE, 2008.

[70] D. A. Reynolds et al. Gaussian mixture models. *Encyclopedia of biometrics*, 741(659-663), 2009.

[71] J. Robinson. The expression and arousal of emotion in music. *The Journal of Aesthetics and Art Criticism*, 52(1):13–22, 1994.

[72] J. A. Russell. Culture and the categorization of emotions. *Psychological bulletin*, 110(3):426, 1991.

[73] S. Schachter and J. Singer. Cognitive, social, and physiological determinants of emotional state. *Psychological Review*, 69(5):379, 1962.

[74] K. R. Scherer, J. Sundberg, B. Fantini, S. Trznadel, and F. Eyben. The expression of emotion in the singing voice: Acoustic patterns in vocal performance. *The Journal of the Acoustical Society of America*, 142(4):1805–1815, 2017.

[75] R. Schluter and H. Ney. Using phase spectrum information for improved speech recognition performance. In *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 01CH37221)*, volume 1, pages 133–136. IEEE, 2001.

[76] H. Schriefers and G. Vigliocco. Speech production, psychology of [repr.]. In *International Encyclopedia of the Social & Behavioral Sciences (2nd ed) Vol. 23*, pages 255–258. Elsevier, 2015.

[77] C. P. Sellors. Carl plantinga and greg m. smith (eds.), Passionate Views: Film, Cognition, and Emotion, 2000.

[78] P. Singh, S. Waldekar, M. Sahidullah, and G. Saha. Analysis of constant-q filterbank based representations for speech emotion recognition. *Digital Signal Processing*, 130:103712, 2022.

[79] V. Stevens. Newman, 1937 stevens ss, volkmann j., newman eb. *A scale for the measurement of the psychological magnitude pitch, Journal of the Acoustical Society of America, JASA*, 8:185–190, 1937.

[80] M. Swain, A. Routray, and P. Kabisatpathy. Databases, features and classifiers for speech emotion recognition: a review. *International Journal of Speech Technology*, 21(1):93–120, 2018.

[81] H. Tak and H. A. Patil. Novel linear frequency residual cepstral features for replay attack detection. In *INTERSPEECH, Hyderabad, India,*, pages 726–730, 2018.

[82] T. Thanapattheerakul, K. Mao, J. Amoranto, and J. H. Chan. Emotion in a century: A review of emotion recognition. In *proceedings of the 10th international conference on advances in information technology*, pages 1–8, 2018.

[83] M. Todisco, H. Delgado, and N. Evans. Constant q cepstral coefficients: A spoofing countermeasure for automatic speaker verification. *Computer Speech & Language*, 45:516–535, 2017.

[84] S. Uthiraa, A. Kachhi, and H. Patil. Linear frequency residual features for infant cry classification. In *submitted in International Conference on Pattern Recognition and Machine Intelligence (PREMI), Kolkata, India*, December 12-15.

[85] S. Uthiraa, A. Pusuluri, and H. Patil. Modified group delay features for emotion recognition. In *submitted in International Conference on Pattern Recognition and Machine Intelligence (PREMI), Kolkata, India*, December 12-15.

[86] S. Uthiraa, B. Singh, and H. Patil. Linear frequency residual features for emotion recognition. In *submitted in International Conference on Pattern Recognition and Machine Intelligence (PREMI), Kolkata, India*, December 12-15.

[87] S. Uthiraa, A. Vora, P. Bonde, A. Pusuluri, and H. A. Patil. Spectral and pitch components of cqt spectrum for emotion recognition. In *rejected in European Signal Processing Conference (EUSIPCO),Helsinki, Finland*, 4-8 September 2023.

[88] S. Uthiraa, A. Vora, and H. Patil. Whisper features for speech emotion recognition. In *submitted in International Conference on Pattern Recognition and Machine Intelligence (PREMI), Kolkata, India*, December 12-15.

[89] E. Vyzas. *Recognition of emotional and cognitive states using physiological data*. PhD thesis, Massachusetts Institute of Technology, USA, 1999.

[90] L. Wang, E. X. Wu, and F. Chen. Contribution of rms-level-based speech segments to target speech decoding under noisy conditions. In *INTERSPEECH*, pages 121–124, 2020.

[91] T. M. Wani, T. S. Gunawan, S. A. A. Qadri, M. Kartiwi, and E. Ambikairajah. A comprehensive review of speech emotion recognition systems. *IEEE Access*, 9:47795–47814, 2021.

[92] Q. Xie, R. K. Ward, and C. A. Laszlo. Automatic assessment of infants' levels-of-distress from the cry signals. *IEEE Transactions on Speech and Audio Processing*, 4(4):253–265, 1996.

[93] S. Yildirim, M. Bulut, C. M. Lee, A. Kazemzadeh, Z. Deng, S. Lee, S. Narayanan, and C. Busso. An acoustic study of emotions expressed in speech. In *Eighth International Conference on Spoken Language Processing*, 2004.

[94] K. Zhou, B. Sisman, R. Liu, and H. Li. Seen and unseen emotional style transfer for voice conversion with a new emotional speech dataset. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 920–924, 2021.

[95] K. Zhou, B. Sisman, R. Liu, and H. Li. Emotional voice conversion: Theory, databases and esd. *Speech Communication*, 137:1–18, 2022.