EMOTION CLASSIFICATION BASED ON GENDER USING TEMPORAL CONVOLUTIONAL NETWORKS AND MULTIPLE FEATURE SETS

A Project Report

Submitted by

Ms. FATHIMA NAZARUDEEN

REG NO : TKM21MEAI04

SEMESTER : IV

In partial fulfillment for the award of the degree of

MASTER OF TECHNOLOGY
IN
Mechanical Engineering (Artificial Intelligence)

Under the guidance of

Prof. SUMOD SUNDAR

Thangal Kunju Musaliar College of Engineering
Kerala

MAY 2023

DECLARATION

I undersigned hereby declare that the project report "EMOTION CLASSIFICATION BASED ON GENDER USING TEMPORAL CONVOLUTIONAL NETWORKS AND MULTIPLE FEATURE SETS", submitted for partial fulfillment of the requirements for the award of degree of Master of Technology of the APJ Abdul Kalam Technological University, Kerala is a bonafide work done by me under supervision of Prof. Sumod Sundar. This submission represents my ideas in my own words and where ideas or words of others have been included, I have adequately and accurately cited and referenced the original sources. I also declare that I have adhered to ethics of academic honesty and integrity and have not misrepresented or fabricated any data or idea or fact or source in my submission. I understand that any violation of the above will be a cause for disciplinary action by the institute and/or the University and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been obtained. This report has not been previously formed the basis for the award of any degree, diploma or similar title of any other University.

Place: Kollam
Date: FATHIMA NAZARUDEEN

Thangal Kunju Musaliar College of Engineering
Centre for Artificial Intelligence

C E R T I F I C A T E

This is to certify that,this report titled ***EMOTION CLASSIFICATION BASED ON GENDER USING TEMPORAL CONVOLUTIONAL NETWORKS AND MULTIPLE FEATURE SETS*** is a bonafide record of the **Project** presented by **FATHIMA NAZARUDEEN (TKM21MEAI04)**, under our guidance and supervision, in partial fulfillment of the requirements for the award of the degree, **M.Tech in Mechanical Engineering (Artificial Intelligence)** in **APJ Abdul Kalam Technological University**.

Project Guide	Project Coordinator	Head of the Department
Prof. Sumod Sundar Assistant Professor Centre for AI	Prof. Chinnu Jacob Assistant Professor Centre for AI	Dr. Imthias Ahamed T P Professor Centre for AI

Internal Examiner

External Examiner

ACKNOWLEDGEMENT

A successful project is a fruitful culmination of efforts by many people, some directly involved and some others indirectly, by providing support and encouragement. Firstly I would like to thank the almighty for giving me the wisdom and grace for making my project a memorable one. I thank him for steering me to the shore of fulfillment under his protective wings.

I express my sincere gratitude to **Dr. T A Shahul Hameed**, Principal of T.K.M College of Engineering for giving me an opportunity to present my project. I would like to thank **Dr. Imthias Ahamed T P**, Professor and Head of the Department, Centre for Artificial Intelligence, TKM College of Engineering, Kollam, for his constant support and encouragement throughout the work.

With a profound sense of gratitude, I would like to express my heartfelt thanks to my guide **Prof. Sumod Sundar** and Project Coordinator, **Prof. Chinnu Jacob**, Assistant Professor, Centre for Artificial Intelligence, TKM College of Engineering, Kollam for their expert guidance, cooperation and immense encouragement. I also extend my thanks to the entire faculty and staff members of the Centre for AI, TKMCE, who have encouraged me throughout this work.

I also express my thanks to my loving parents and friends, for their support and encouragement in the successful completion of this work.

Fathima Nazarudeen

Abstract

Emotion recognition from speech has become a significant research area due to its potential applications in diverse fields such as mental health, human-robot interaction, and virtual assistants. The work presents an approach to classify emotions based on gender into 14 different classes by concatenating four emotional speech datasets: RAVDESS, SAVEE, TESS, and CREMA-D. Multiple features, including MFCCs, spectral contrast, pitch, centroid, roll-off, onset flux, entropy, and ZCR, are extracted from the audio files, and these features are utilized as inputs to a Temporal Convolutional Network (TCN) for emotion classification. TCN is trained to learn high-level features from the extracted features of the four datasets, and these features are then utilized for the classification of emotions. In this approach, different combinations of feature sets and the TCN classifier are evaluated to identify the optimal combination that achieves the highest training and validation accuracies. The combination of MFCC and entropy achieved the highest accuracy rates, with 99.4% and 99.01% for training and validation, respectively. The other combinations of features also achieved high accuracy rates, with some variations in performance. The proposed methodology is effective in accurately classifying emotions from speech, and the use of TCN in conjunction with a variety of feature sets, including the extraction of high-level features, significantly improved the performance of the emotion classification system.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Speech emotion recognition is an emerging field that has gained a lot of attention in recent times due to its applications in various domains, including mental health therapy, marketing, and education. The ability to analyze emotional cues from speech signals provides valuable insights into human behavior and has the potential to advance the development of more empathetic and effective interventions, products, and services [1]. One important aspect of speech emotion recognition research is the consideration of gender-based differences in emotional expression. Understanding these differences can improve the precision of speech emotion recognition systems and assist in better recognizing emotional expressions in both men and women. This, in turn, can facilitate the creation of personalized mental health therapies and emotion-aware human-computer interfaces. Furthermore, utilizing speech emotion recognition technology in marketing can aid organizations in developing more effective marketing and advertising campaigns by gaining insights into client behavior and preferences. Similarly, in education, speech emotion recognition helps teachers provide tailored feedback and support to students based on their emotional state.

Koduru et al. [2] proposed a method by using various feature extraction algorithms, noise removal, global feature selection, and machine learning classification to identify emotions from features including MFCC, DWT, pitch, energy, and ZCR. Chen et al. [3] proposed a three-dimensional attention-based convolutional recurrent neural networks which involves processing Mel-spectrograms with deltas and delta-deltas to reduce emotional irrelevant factors from the speech.

In this work, the main objective is to accurately classify emotions from audio files based on gender into 14 different classes. To achieve this, a number of features are extracted from the audio files including MFCCs, spectral contrast, pitch, spectral centroid, spectral roll-off, onset flux, spectral flatness entropy, and zero crossing rate (ZCR). These features provide comprehensive information about the spectral and temporal characteristics of the audio signal. The extracted features are then fed to a Temporal Convolutional Network (TCN) for classification. The TCN model was utilized to learn additional high-level features that capture long-term dependencies in the audio signal which aims to improve the efficiency of the emotion classification model. TCNs are a type of neural network that is particularly well-suited for processing time series data, as they can capture complex temporal patterns over long time spans. The TCN model in this work utilized convolutional layers with dilated

convolutions and a dense layer with softmax activation. The dilated convolutions enable the model to capture temporal patterns at different scales, while the dense layer with softmax activation allows the model to output a probability distribution over the 14 emotion classes for a given input audio signal.

The contributions of the work are as follows:

- Classify emotions from audio files based on gender.
- Features such as MFCCs, spectral contrast features, pitch, spectral centroid, spectral roll-off, onset flux, spectral flatness entropy, and zero crossing rate (ZCR) are extracted from the audio files.
- TCNs are used to learn high-level features that can depict or represent long-term dependencies in the audio signal. The TCN model consists of convolutional layers with dilated convolutions and a dense layer with softmax activation. The dilated convolutions enable the model to capture temporal patterns at different scales.

Chapter 2

Literature Survey

In the literature, numerous researchers have tackled the problem of Speech Emotion Recognition (SER). This section summarises some of these research endeavors and highlights their accomplishments.

The framework proposed by Dias Issa et al. [4] for speech emotion recognition is a deep learning-based approach that utilizes a one-dimensional CNN architecture. The CNN is trained to classify emotions using a combination of five different audio features, namely the mel-scale spectrogram, spectral contrast, Tonnetz representation, MFCC, and chromagram. These features are extracted from the audio recordings and fed as inputs to the CNN. To improve the performance of the proposed framework, it was suggested incorporating an auxiliary neural network to extract high-level features, which can be used as input to the CNN. In addition, more data augmentation techniques can be applied to further enhance the model's ability to generalize to unseen data. Another way to improve the accuracy of the model is by adding more LSTM layers, which can help capture temporal dependencies in the data. It is important to note that the order in which the audio features are stacked can also impact the accuracy of the model. Therefore, altering the feature order may result in different classification accuracies.

Jiang et al. [5] proposed a method for speech emotion recognition using a combination of Deep Convolutional Neural Network and Simple Recurrent Unit. The proposed method utilizes log Mel-spectrograms, which are a commonly used representation of audio signals in speech processing, and are known to provide a compact and efficient representation of the audio signal. It extracts log Mel-spectrograms with static, delta, and delta-delta acoustic features.These spectrograms are then segmented into fixed-size frames and fed into a pre-trained AlexNet for feature extraction. The output features of the AlexNet are then used as inputs to an SRU model that aggregates segment-level features and maps them to a fixed-size representation. Finally, the fixed-size representation is fed into a SoftMax classifier to predict the emotion. One of the limitations of this approach is that the performance of the model may be affected by the segment size used as input. The authors chose a fixed segment size of 2 seconds, but this may not be the optimal size for all types of speech signals. Additionally, the use of log Mel-spectrograms may not capture all aspects of emotional features present in the speech signal. For example, prosody and intonation may also play an important role in conveying emotional information that is not fully captured by log Mel-spectrograms.

Farooq et al. [6] proposed a method for speech emotion recognition (SER) that utilizes a deep convolutional neural network (DCNN) for feature extraction and several classification algorithms for emotion classification. The aim of the study was to explore the benefits of using a DCNN for SER and to examine the effectiveness of different classification algorithms for this task. The study used four publicly available datasets for emotional speech: the Berlin Dataset of Emotional Speech (Emo-DB), Surrey Audio Visual Expressed Emotion (SAVEE), Interactive Emotional Dyadic Motion Capture (IEMOCAP), and the Ryerson Audio Visual Dataset of Emotional Speech and Song (RAVDESS). However, limitations included the need for large labeled data and further research is needed to explore gender effects on speech emotion signals.

Aggarwal et al. [7] proposed a novel model for recognizing speech emotions that utilizes two-way feature extraction and deep transfer learning. The model extracts two sets of features from the speech data, one using superconvergence and the other using a pre-trained VGG-16 model. The first set of features is obtained by applying superconvergence, a technique that accelerates the convergence of deep neural networks by optimizing learning rate and momentum values. The second set of features is obtained by feeding the speech data through a pre-trained VGG-16 model, which has been widely used in image recognition tasks and has shown promising results in speech recognition as well. To further enhance the discriminative power of the extracted features, Principal Component Analysis (PCA) is applied to the first set of features. PCA is a statistical technique that transforms a set of correlated variables into a new set of uncorrelated variables, called principal components, which retain most of the original information while reducing the dimensionality of the data. This step helps to remove redundant and irrelevant information from the features, thereby improving the classification accuracy. The proposed model is evaluated on the RAVDESS dataset, which contains speech recordings of ten different actors, portraying seven basic emotions. However, the study's limitations include the RAVDESS dataset used only consisting of North American speakers and people of median age, suggesting the proposed model should be applied to other datasets in the future to evaluate its effectiveness across different demographics.

Mehmet Bilal Er [8] proposed a hybrid architecture based on acoustic and deep features to improve the accuracy of speech emotion recognition. The method consists of three stages: feature extraction, feature selection, and classification. Acoustic features such as Root Mean Square energy (RMS), Mel-Frequency Cepstral Coefficients (MFCC), and Zero-crossing Rate are extracted from voice recordings. Spectrogram images of the original sound signals are then fed into pre-trained deep neural networks including VGG16, ResNet18, ResNet50, ResNet101, SqueezeNet, and DenseNet201 to extract deep features. A hybrid feature vector is created by combining the acoustic and deep features, and the ReliefF algorithm is used for feature selection. Finally, a linear Support Vector Machine (SVM) is used for classification. The method is being evaluated on three popular datasets, RAVDESS, EMO-DB, and IEMOCAP. However, the use of pre-trained deep neural networks and linear SVM for feature extraction and classification respectively may limit the performance of the proposed method. Furthermore, the generalizability of the proposed method to other datasets or real-world scenarios may need to be further explored.

Mustaqeem et al. [9] proposed an approach for Speech Emotion Recognition (SER) that involves selecting key sequence segments using Radial Basis Function Network (RBFN) similarity measurement in clusters, converting them into spectrograms with Short Time Fourier Transform (STFT), and extracting discriminative features with a Convolutional Neural Network (CNN) model, which are then normalized and fed to a deep Bidirectional Long Short-Term Memory (BiLSTM) for recognizing the final state of emotion. The proposed approach aims to capture both the global and local features of speech signals to improve the accuracy of emotion recognition. However, one potential limitation of this approach is that the RBFN clustering algorithm used for key sequence selection may not always select the most relevant segments for emotional state recognition, leading to reduced accuracy. The performance of the proposed approach may also be affected by the choice of hyperparameters in the CNN and BiLSTM models. Additionally, the generalizability of the approach to different datasets or real-world scenarios needs to be further evaluated.

Chapter 3

Methodology

3.1 Objective

To accurately identify the high-level features and classify the speaker's emotional state based on their gender using efficient deep learning techniques.

3.2 Techniques Used

The proposed methodology for emotion classification based on gender as shown in Fig. 3.1 starts with concatenating four different datasets containing emotional speech data, namely RAVDESS [10], SAVEE [11], TESS [12], AND CREMA-D [13]. This step is crucial to enhance the diversity of the training data and improve the ability of the proposed emotion classification model to generalize. The audio files are then pre-processed, and a set of acoustic features are extracted, including MFCCs, spectral contrast features, pitch, spectral centroid, spectral roll-off, onset flux, spectral flatness entropy, and zero crossing rate(ZCR). These features are known to capture relevant information about the emotional content of speech signals.

In order to enhance its efficacy, the extracted features were then fed into a **TCN model**, which is designed to capture long-term dependencies within sequential data. This allowed the TCN to acquire high-level representations of the audio signals that capture the complex relationships between different acoustic features and emotional content. This step is essential for building a robust and accurate emotion classification model.

Finally, the learned high-level features are used to classify emotions based on gender. The task of classification is structured as a multi-class problem, with the model utilizing the input features and gender of the speaker to predict the corresponding emotional label. The proposed approach leverages the power of deep learning models, such as TCNs, to capture the complex relationships between acoustic features and emotions and model gender-specific differences in emotional expression. The methodology has the potential to improve the accuracy of emotion recognition by modeling gender-specific differences and can be applied in various fields, such as affective computing, mental health assessment, and human-computer interaction.

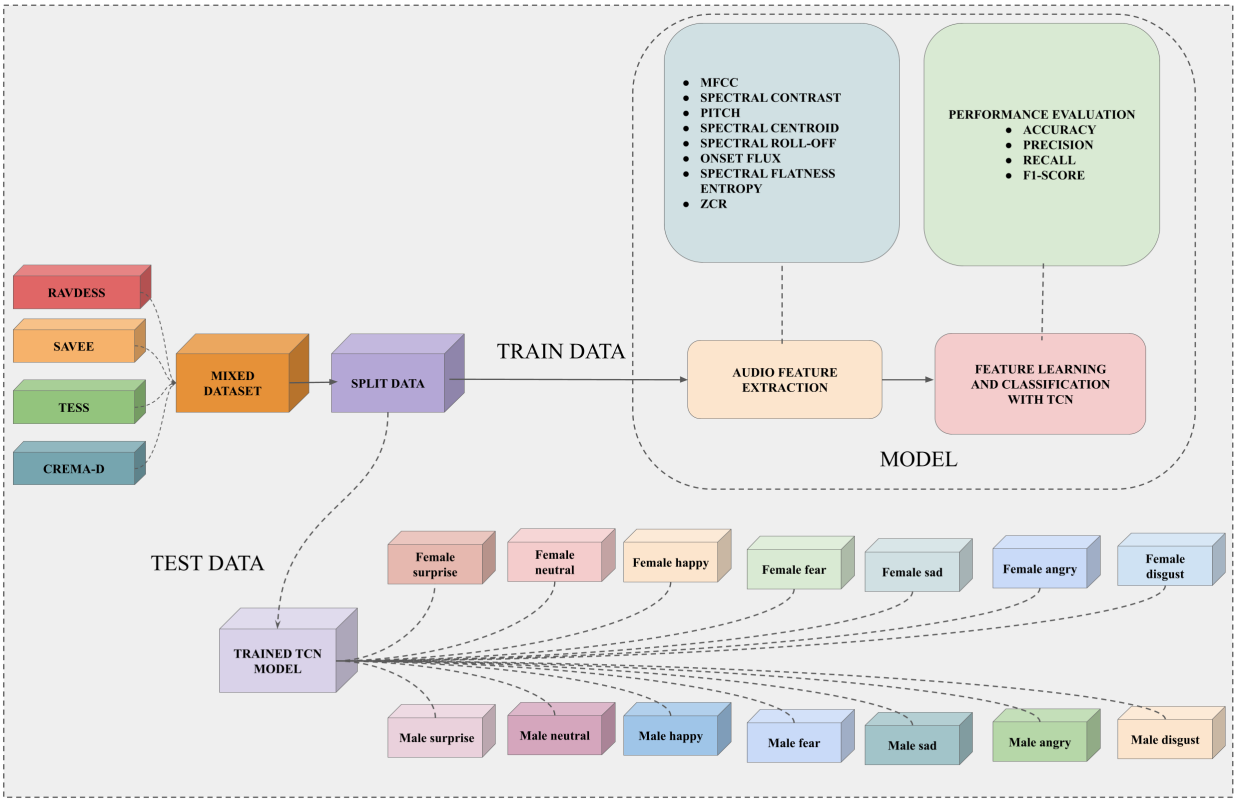

Figure 3.1: Methodology

3.2.1 Feature Extraction

Feature extraction is an essential step in speech emotion recognition, which involves extracting relevant information from audio signals that can be used to classify emotions. Librosa is a popular Python library for audio processing, which provides a range of functions for extracting various audio features that are relevant to speech emotion recognition.

The advantage of using librosa for speech emotion recognition is that it provides a convenient and efficient way to extract multiple features from audio signals. This can help improve the accuracy and robustness of emotion classification models, as different features capture different aspects of the audio signal. The features extracted using librosa include MFCCs, spectral contrast features, pitch, spectral centroid, spectral roll-off, onset flux, spectral flatness entropy, and zero crossing rate(ZCR).

- **MFCC** (Mel-frequency Cepstral Coefficients)

  MFCC [14] is a feature extraction method used in speech and audio processing. Block diagram of the MFCC extraction process is shown in Fig. 3.2. The resulting MFCCs are a set of coefficients that capture the spectral shape of the audio signal and are often used as features for speech and audio processing tasks such as speech recognition, speaker recognition, and emotion recognition. Here 40 MFCC features along with their first and second order derivatives are extracted.

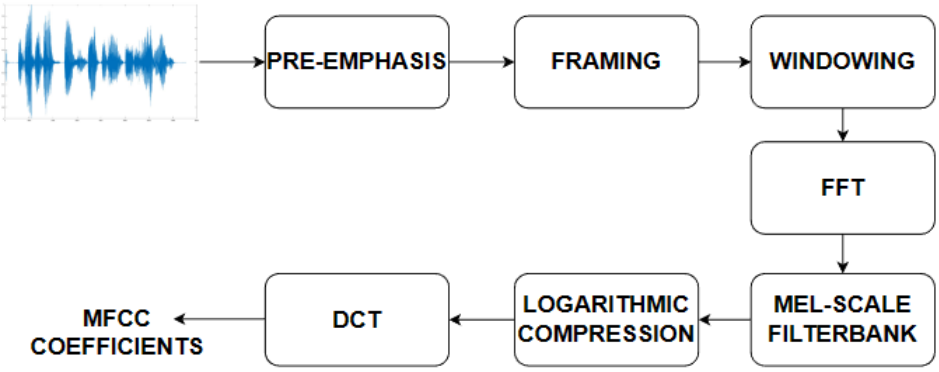

Figure 3.2: MFCC

The steps to extract the MFCC coefficients is as follows:
Pre-Emphasis: The first step is to apply a Pre-Emphasis filter to the audio signal to boost the higher frequencies and compensate for the natural attenuation of the lower frequencies during recording and transmission.

Framing: The audio signal is then divided into small frames of equal duration.

Windowing: Each frame is then multiplied by a window function, such as the hamming window, to reduce the spectral leakage and smoother the edges of the frame.

Fourier Transform: It is applied to each frame to convert it from the time domain to the frequency domain.

Mel Filterbank: The power spectrum obtained from the Fourier Transform is then passed through a series of triangular filters spaced on the Mel scale, that mimics the human ear's response to different.

Logarithm/ Log Compression: The output of each filter is then compressed using a logarithmic function to emphasize the lower frequencies and reduce the effect of higher frequencies. The logarithmic scale represents the human auditory system's response to loudness.

Discrete Cosine Transform (DCT): Applies DCT to the logarithmic Mel-scale spectrum to extract the most significant coefficients.

The 40 MFCC coefficients are:
MFCC 1 : The first coefficient represents the overall energy of the signal. It is often called cepstral mean or constant coefficient because it captures the average energy across all frequency bands. This coefficient is calculated by taking the logarithm of the

energy in each frequency band and then averaging across all bands. This process helps to normalize the signal and removes any differences in overall energy between different signals.

MFCC 2-13 : capture the shape of the power spectrum, which is also known as the spectral envelope. These coefficients tell how much energy is present in different frequency bands. They are calculated by applying a filter bank to the power spectrum and then taking the logarithm of the energy in each band. The filter bank is designed to approximate the non-linear frequency response of the human ear, which is why it is often referred to as the mel-frequency domain.

MFCC 14-19 : Represent the first-order derivatives of the spectral envelope coefficients. They capture how the spectral envelope changes over time. These coefficients are calculated by taking the derivative of each spectral envelope coefficient with respect to time. This process helps to capture how the frequency distribution of a sound evolves over time and can be used to distinguish between different types of sounds.

MFCC 20-27 : Represent the second-order derivatives of the spectral envelope coefficients. They tell about how the shape of the sound is changing over time i.e. whether it is getting sharper or flatter.

MFCC 28 : Represents the total energy in the high-frequency band of the audio signal. This coefficient is calculated by summing the energy in all frequency bands above a certain cutoff frequency. This coefficient is useful for distinguishing between sounds that have similar spectral characteristics but differ in their high-frequency content.

MFCC 29-33 : Represent the first-order derivatives of the high-frequency band energy coefficients. These coefficients capture how the high-frequency energy changes over time. These coefficients are calculated by taking the derivative of the high-frequency energy coefficient with respect to time. This process provides information on changes in emotional intensity or other important characteristics of the sound.

MFCC 34-40 : Represent the second-order derivatives of the high-frequency band energy coefficients. They capture how the curvature of the high-frequency energy is changing over time. These coefficients are calculated by taking the second derivative of the high-frequency energy coefficient with respect to time. This process provides additional information on the emotional content of the audio signal and can be useful for speech recognition or music analysis applications.

The MFCC (Mel-Frequency Cepstral Coefficients) of speech samples from four different speakers, including two male and two female, expressing two different emotions, were visualized in this work. The first sample was from a female speaker expressing anger in Fig. 3.3, the second from a male speaker also expressing anger in Fig. 3.4, the third sample from a female speaker expressing happiness in Fig. 3.5, and the fourth from a

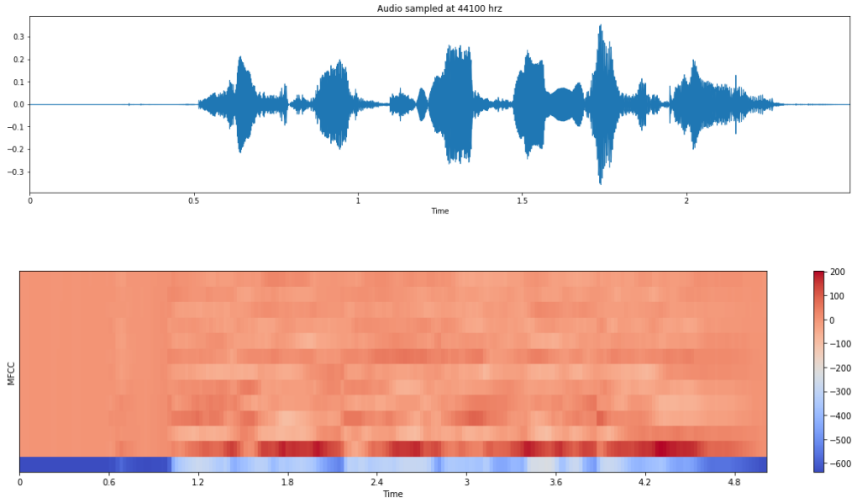

Figure 3.3: MFCC of female speaker expressing anger

male speaker expressing happiness in Fig. 3.6. The MFCC features were extracted from each sample and presented using a heat map. This analysis can provide insights into the acoustic characteristics of speech associated with different genders and emotions. From the MFCC plots, it can be observed that the first band at the bottom is the most distinctive band over the other bands. Since the time window is a short one, the changes observed overtime does not vary greatly. The key feature here is capturing the information contained in the various bands.

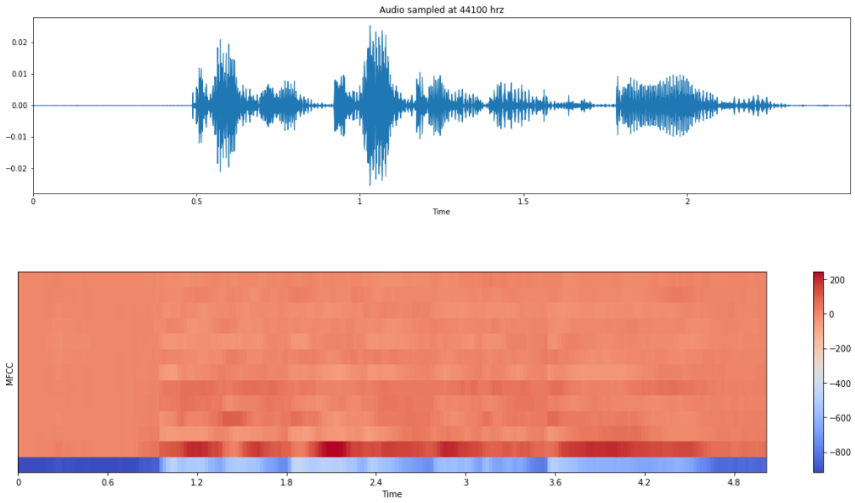

Figure 3.4: MFCC of male speaker expressing anger

The Mel-Frequency Cepstral Coefficients (MFCC) of audio samples from the RAVDESS dataset were analyzed to investigate differences in acoustic characteristics between male and female speakers. Line plots were generated to compare the MFCC feature vectors of two audio samples, one from a male speaker and another from a female speaker expressing anger, as well as two audio samples expressing happiness.

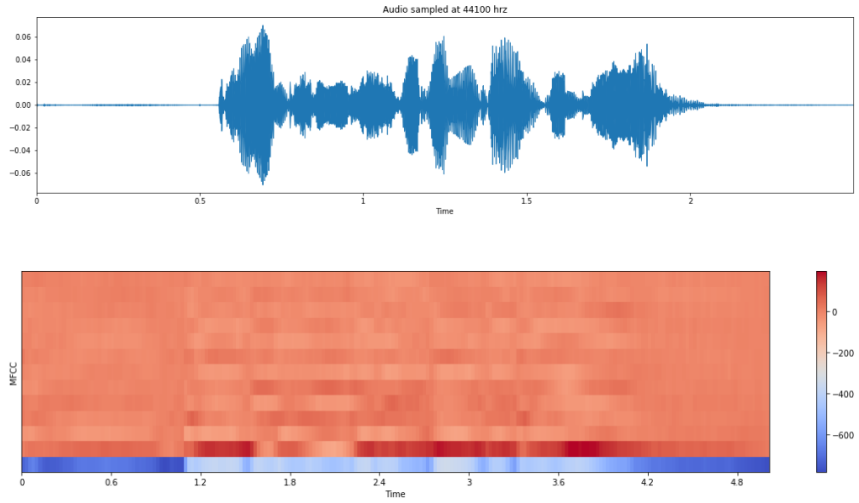

Figure 3.5: MFCC of female speaker expressing happiness

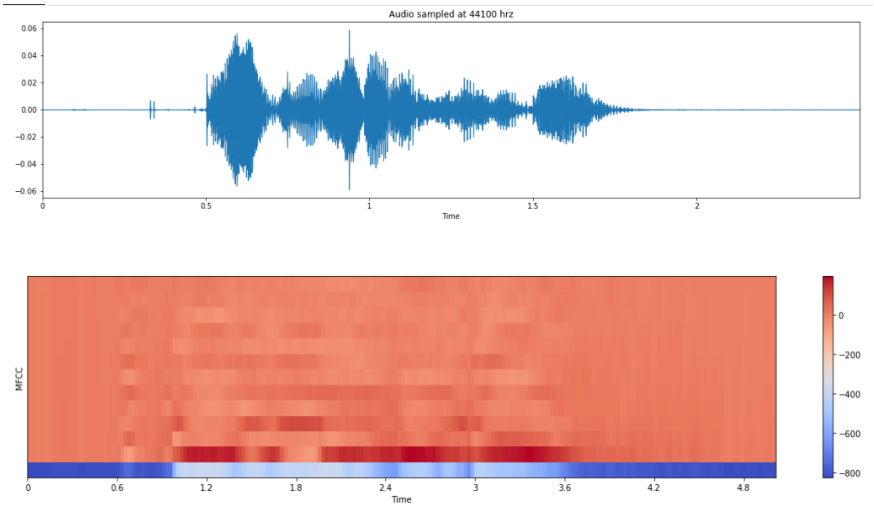

Figure 3.6: MFCC of male speaker expressing happiness

In each case, the librosa library in Python was used to extract the MFCC features, and the mean of these features across all frames was computed to obtain representative MFCC feature vectors for the audio samples. The line plots indicated that male and female speakers exhibited distinct acoustic characteristics, with differences observed in the distribution of MFCC values across frequency ranges.

For the angry samples, the female speaker had higher MFCC values in the middle frequency range, while the male speaker had higher values at lower and higher frequency ranges as shown in Fig. 3.7. For the happy samples, the female speaker had higher MFCC values in the mid to high frequency range, while the male speaker had higher values at lower frequency ranges as shown in Fig. 3.8. These findings suggest that male and female speakers may exhibit distinct acoustic characteristics even when expressing the same emotion.

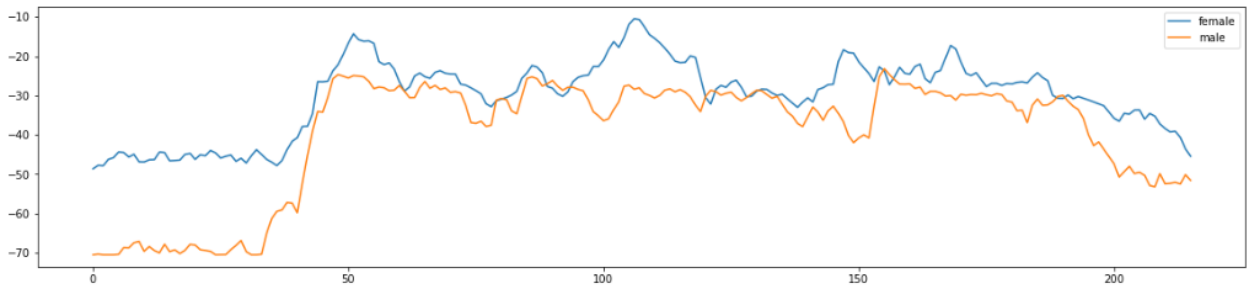

Figure 3.7: Comparison of Mel-Frequency Cepstral Coefficients between Male and Female Speakers Expressing Anger

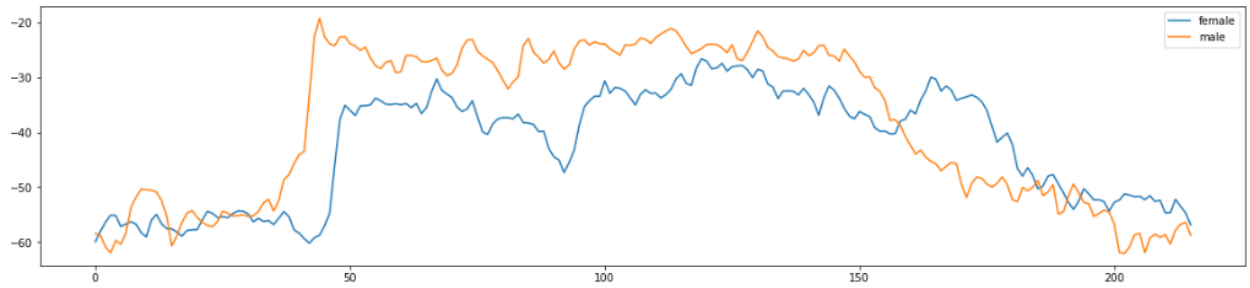

Figure 3.8: Comparison of Mel-Frequency Cepstral Coefficients between Male and Female Speakers Expressing happiness

- **Spectral contrast**

  Spectral contrast is a feature that can provide useful information about the spectral content of a signal, which can be helpful in emotion recognition. It measures the relative difference between energy in different frequency bands and can capture important perceptual features of the signal, such as the harmonic structure of the voice or the presence of formants. In speech processing, spectral contrast can be used to distinguish between different emotional states by analyzing the spectral changes that occur when a person is expressing different emotions. For example, when a person is happy, their voice may have higher energy in the higher frequency bands, which can be detected using spectral contrast analysis. Conversely, when a person is sad or depressed, their voice may have a lower energy in the higher frequency bands. It is typically computed by dividing the frequency spectrum into several frequency bands, and then comparing the energy in each band to the energy in adjacent bands. Moreover, spectral contrast can be used in combination with other features, such as pitch and intensity, to improve the accuracy of emotion recognition systems. By combining different features, it is possible to capture the complex and dynamic nature of emotions and provide a more accurate and nuanced representation of the emotional state of the speaker [15].

- **Pitch**

  Pitch is a term used to describe the response of the human ear to the frequency of a sound wave. It is closely related to the frequency of the sound wave, with higher

frequencies corresponding to higher pitch and lower frequencies corresponding to lower pitch. Pitch can be a useful feature for emotion recognition and can also provide information about the gender of the speaker. Typically, females have higher pitch than males, and analyzing the pitch of a signal can help to determine the gender of the speaker [2].

The benefit of pitch in emotion recognition is that it can help to detect emotional states that are not easily observable from other modalities, such as facial expressions or body language. For example, a person may try to conceal their emotional state by controlling their facial expressions or body posture, but their voice may still reveal their emotional state through variations in pitch.

- **Spectral centroid**

  Spectral centroid is a feature of audio signals that provides information about the frequency content of the signal. It measures where most of the energy of the audio signal is concentrated in terms of frequency. It is calculated by finding the weighted average of all the frequencies present in the signal, where the weights are the amplitudes of the spectral components. In other words, it is a measure of the center of mass of the spectral distribution of an audio signal. It is used to analyze changes in spectral content of speech over time.

  Changes in spectral centroid can indicate changes in the "brightness" or "darkness" of the sound. For example, when speaker is angry the spectral centroid of their speech tends to shift towards higher frequencies indicating a brighter, more intense sound. Conversely, when a speaker becomes sad, the spectral centroid of their speech tends to shift towards lower frequencies indicating a darker, more deeper sound. Spectral centroid of male speech tends to be lower than that of female speech. So it can be used as a feature to differentiate between male and female speakers.

- **Spectral roll-off**

  Spectral roll-off is a measure of the steepness of the spectral slope of an audio signal. It provides information about how quickly the spectral energy diminishes as the frequency increases. In speech processing, it can be used to analyze the spectral content of the voice and identify gender differences. Typically, female voices have a higher spectral roll-off than male voices, which means that their spectral energy decreases more quickly as the frequency increases. This difference in spectral roll-off can be used to distinguish between male and female speakers, which can be useful in several applications, including speaker identification and emotion recognition.

- **Onset flux**

  Onset flux is a measure of the rate of change of the signal at its onset. It provides information about the spectral content and temporal shape of the onset. Extracting onset

flux can be relevant for emotion recognition based on gender, as it captures gender-related differences in vocal dynamics and emotional expressiveness at the start of an utterance. Female speakers have a higher onset flux than male speakers, which means that their voice changes more rapidly at the start of an utterance. This difference in onset flux can be used to distinguish between male and female speakers.

- **Spectral flatness entropy**

  Spectral flatness entropy provides information about the spectral flatness of an audio signal. Spectral flatness is the measure of how evenly the energy of an audio signal is distributed across different frequencies in the spectrum. It is calculated by taking the ratio of the geometric mean of the power spectrum to the arithmetic mean of the power spectrum of the signal. A high value of this feature indicates a more tonal signal, while a low value indicates a more noisy signal. It captures gender-related differences in the spectral balance which is the distribution of energy across different frequency bands in an audio signal. Male speech signals generally have a higher energy in the low-frequency range, while female speech signals have a more even distribution of energy across the frequency spectrum.

- **Zero crossing rate**

  The zero crossing rate can be defined as the rate at which a signal transitions from a positive value to zero, or from zero to a negative value (and vice versa). It provides information about the spectral characteristics of the signal. ZCR is a robust feature that can be used in noisy environments where other features may ne affected. Since ZCR is based on the sign changes of the audio signal, it is less affected by the noise than other features that rely on the amplitude of the signal. Male speech signals have a higher ZCR than female speech signals due to differences in the vocal tract structure [16].

3.2.2 Temporal Convolutional Network(TCN)

Time Convolutional Networks (TCNs) have emerged as a popular approach for speech emotion recognition, owing to their ability to capture the temporal dependencies inherent in audio data. This is achieved by first extracting relevant features from the audio signal, which are then input to a TCN model consisting of a series of dilated convolutional layers followed by one or more fully connected layers. The convolutional layers capture local patterns in the input features, while the dilated convolutional layers enable the model to capture long-term temporal dependencies in the audio signal. The resulting trained TCN model can then be utilized to predict the emotion label of new audio data with high accuracy. Furthermore, the high-level features extracted by the TCN model can be employed for further analysis and visualization of the emotion data, providing deeper insights into the underlying patterns and trends in the data.

TCNs are a popular variation of Convolutional Neural Networks (CNNs) for sequence modeling tasks. They combine the benefits of Recurrent Neural Networks (RNNs) and CNNs by utilizing multiple layers of dilated convolutions and padding techniques to handle varying sequence lengths and identify dependencies between items that are not adjacent. Unlike traditional CNNs, TCNs employ causal convolutions as shown in Fig. 3.9, where the output at a given time step is only convolved with elements from that step and earlier in the previous layer, thereby preventing information leakage from future to past.

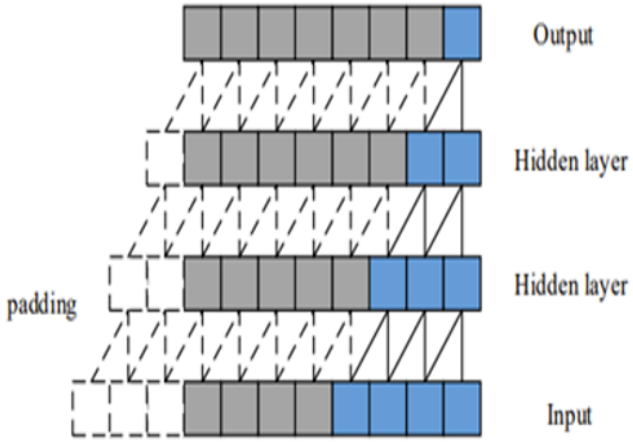

Figure 3.9: Causal convolution with filter kernel size k=2

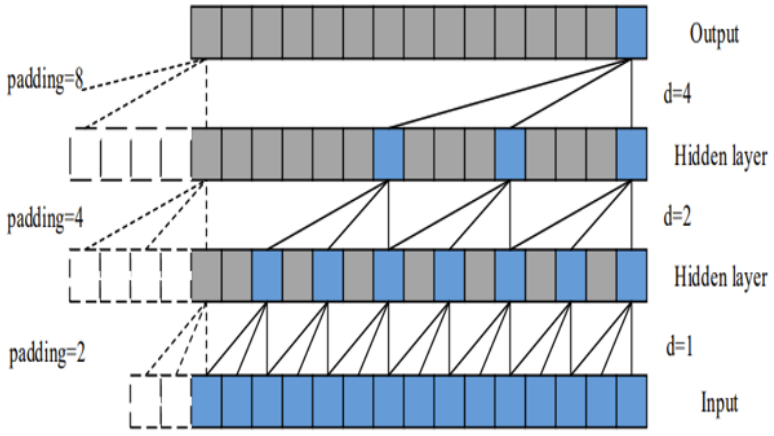

Figure 3.10: Dilated causal convolution with dilated factors d=1, 2, 4 and filter size k=3

However, the use of causal convolutions limits the receptive field size unless a large number of layers are stacked, leading to significant computational overhead. To overcome this limitation, TCNs utilize dilated convolutions as shown in Fig. 3.10, which enable exponentially larger receptive fields without requiring a significant increase in

the number of layers.

In dilated convolutions, the filter is applied over a region larger than its size by skipping input values with a given step, referred to as the dilation rate. In TCNs, the dilation rate is commonly increased exponentially with the depth of the network, allowing the receptive field to cover each input in the history and providing an extremely long effective history size.

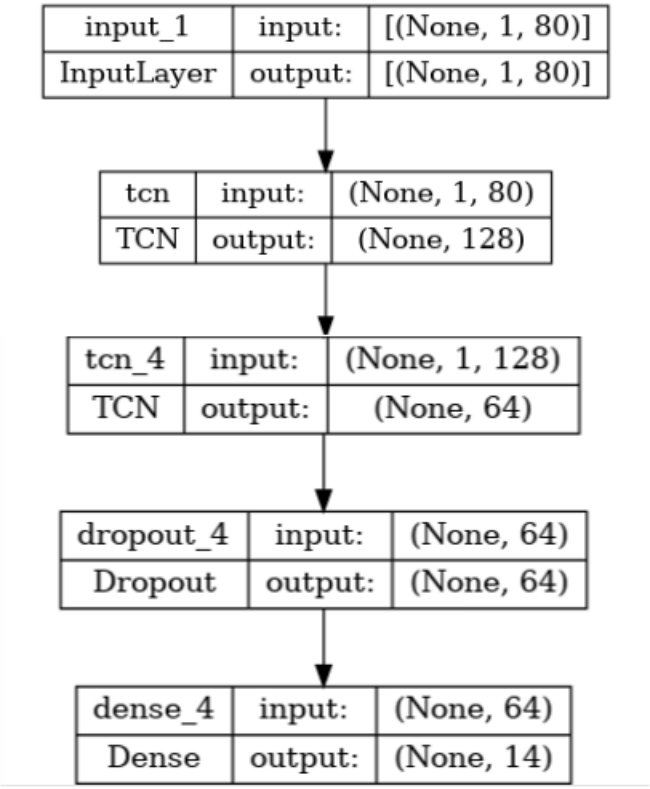

Figure 3.11: TCN

The TCN model used in this work consists of a two-layered TCN as shown in Fig. 3.11, where the first layer comprises 128 filters and a kernel size of 6. This layer functions as the primary feature extraction stage and employs a dilations parameter of [1,2,4,8,16,32,64,128], resulting in an exponentially growing receptive field. This configuration enables the model to effectively capture long-term temporal dependencies in the audio signal by processing progressively larger temporal windows. The extracted features from the first TCN layer are then passed to the second TCN layer, which utilizes a dilations parameter of [1,2,4,8,16,32,64] and 64 filters. The subsequent layer is a dense layer with 14 units and a softmax activation function, enabling the model to accurately classify the input audio signal into one of the 14 classes.

The model is compiled with the Adam optimizer and employs the categorical cross-entropy loss function. By employing the two-layered TCN architecture with an increas-

ing dilation parameter, the model can learn complex and high-level features, allowing it to achieve superior performance in capturing long-term temporal dependencies in the audio signal. The subsequent dense layer with softmax activation further enhances the model's classification accuracy.

3.2.3 Datasets

The proposed methodology utilizes four distinct audio datasets, namely RAVDESS, SAVEE, TESS, and CREMA-D, which are commonly used by researchers in the field of emotion recognition.

RAVDESS

The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) dataset [10] contains 7356 files, including audio and video recordings of actors performing emotional speech and song. Of these files, 1440 are audio files, each with a duration of approximately 3-5 seconds. The audio files are organized into 24 different actors (12 male and 12 female) performing in 8 different emotional categories: calm, happy, sad, angry, fearful, surprise, disgust, and neutral. The age range of the actors is 20-30 years old for the younger group (10 males and 10 females) and 50-60 years old for the older group (2 males and 2 females).

SAVEE

The Surrey Audio-Visual Expressed Emotion (SAVEE) dataset [11] contains a total of 480 audio files, each with a duration of approximately 3-5 seconds. The dataset includes recordings of 4 male actors (aged between 26-46 years) performing in 7 different emotional categories: anger, disgust, fear, happiness, sadness, surprise, and neutral.

TESS

The Toronto Emotional Speech Set (TESS) dataset [12] includes a total of 2800 audio files, with each file being approximately 3-5 seconds long. The dataset comprises recordings of four professional actors (two male and two female), with each actor performing in seven different emotional categories: anger, disgust, fear, happiness, pleasant surprise, sadness, and neutral. While the age of the actors in the TESS dataset is not explicitly stated, it is known that they are all adults.

CREMA-D

The Crowd-Sourced Emotional Multimodal Actors Dataset (CREMA-D) [13] consists of a collection of 7,442 unique audio clips sourced from a diverse group of 91 actors, 48 of whom are male and 43 of whom are female. The actors range in age from 20 to 74 and represent various racial and ethnic backgrounds, including African American,

Asian, Caucasian, Hispanic, and Unspecified. The audio clips in the dataset correspond to six basic emotions: Anger, Disgust, Fear, Happy, Neutral, and Sad.

3.2.4 Performance Evaluation

To evaluate the performance of Speech emotion recognition models (SER), several metrics are used, including accuracy, precision, recall, and F1 score. These metrics provide a quantitative measure of the model's ability to recognize emotions in speech signals. By using these metrics, it will be able to compare different SER models and choose the one that performs the best.

Accuracy is a metric that indicates how often the model correctly predicts the emotional state of a speaker. It is calculated as the ratio of the number of correctly predicted emotions (true positives and true negatives) to the total number of emotions in the dataset. An accurate model will have a high accuracy score, indicating that it can correctly recognize emotions most of the time. The accuracy score can be calculated as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.1}$$

where TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives, and FN is the number of false negatives.

Precision is a metric that indicates how often the model correctly predicts a positive emotion out of all the emotions it classified as positive. It is calculated as the ratio of the number of true positives to the sum of true positives and false positives. A model with high precision is one that has a low false positive rate, meaning that it does not wrongly classify negative emotions as positive. Precision can be calculated as:

$$Precision = \frac{TP}{TP + FP} \tag{3.2}$$

where TP is the number of true positives and FP is the number of false positives.

Recall is a metric that indicates how well the model correctly identifies positive emotions in the dataset. It is calculated as the ratio of the number of true positives to the sum of true positives and false negatives. A model with high recall is one that can identify most of the positive emotions in the dataset. Recall can be calculated as:

$$Recall = \frac{TP}{TP + FN} \tag{3.3}$$

where TP is the number of true positives and FN is the number of false negatives.

F1 Score is a combination of precision and recall that provides a balance between the two metrics. It is useful when the dataset has an imbalanced class distribution, where one class is much more prevalent than the other. The F1 score considers both false positives and false negatives and penalizes models that have either a high false positive or false negative rate. It is a value between 0 and 1, where a higher score indicates a better model performance. The F1 score can be calculated as:

$$F1\ Score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

Chapter 4

Results and Discussion

The suggested model is implemented in Python, along with the essential libraries Librosa, NumPy, Pandas, Scikit-learn, and other. The four datasets were concatenated and relevant features, including MFCCs, spectral contrast features, pitch, spectral centroid, spectral roll-off, onset strength, spectral flatness entropy, and ZCR were extracted. These features were then utilized as inputs to train the Temporal Convolutional Network (TCN) model to learn and identify high-level features for the accurate classification of emotions into 14 different classes. Table 4.1 displays the effectiveness

Table 4.1: Classification performance of TCN with different feature combinations

No:	Features Extracted	Training Accuracy (%)	Validation Accuracy (%)
1.	Spectral Contrast	96.93	93.63
2.	MFCC+ZCR	98.65	96.92
3.	MFCC+Pitch	98.92	97.70
4.	MFCC+Spectral Contrast	98.46	97.94
5.	MFCC+Centroid	98.89	98.03
6.	MFCC	98.97	98.19
7.	MFCC+Spectral Contrast+Pitch+ Centroid+Roll-off+ Onset Flux+ Entropy+ZCR	98.96	98.36
8.	MFCC+Roll-off	98.68	98.64
9.	MFCC+Flux	98.96	98.85
10.	**MFCC+Entropy**	**99.40**	**99.01**

of the TCN model with various feature combinations for audio classification. The proposed approach involves extracting different combinations of features from the audio signals, which are then fed into the TCN model to learn high-level representations. The combination of MFCC and entropy achieved the highest training and validation accu-

racies, with 99.4% and 99.01%, respectively. MFCC is a widely used feature in audio processing that captures the spectral shape of the audio signal. In contrast, entropy measures the amount of uncertainty or randomness in the signal. By combining these two features, the model can capture both the spectral and temporal characteristics of the audio signals, which are critical for distinguishing between different categories of audio. The high accuracies obtained suggest that the proposed approach of using TCN is a promising method for audio classification tasks.

The choice of feature representation can significantly impact the accuracy of an audio classification model. From the accuracy plots in Fig. 4.1, it can be observed that the accuracy of the TCN model with spectral contrast features is lower than the other feature combinations. This is because spectral contrast features are more sensitive to variations in recording conditions, such as background noise or reverberation, which can affect the spectral envelope of the sound. Other feature combinations, such as MFCC and entropy, are less sensitive to such variations and perform better in this task. The high accuracies achieved by the combination of MFCC and entropy, with 99.4% training accuracy and 99.01% validation accuracy, respectively, suggest that these features are highly effective in accurately classifying emotions. However, other feature combinations also achieved high accuracy rates, with some variations in performance depending on the specific combination of features used. For example, some feature combinations may be more effective in capturing certain aspects of the speech signal that are important for emotion recognition, while others may be less effective.

Fig. 4.2 shows that MFCC+Entropy has the lowest training loss (0.0252) and validation loss (0.0403), indicating good model performance and generalization. Spectral Contrast has the highest training loss (0.1047) and validation loss (0.2531), suggesting lower performance compared to other feature combinations.

Fig. 4.3 and Fig. 4.4 shows the confusion matrix based on the performance of TCN with different feature combinations. The rows represent the actual emotions, while the columns represent the predicted emotions. Each cell in the matrix shows the number of times that the classifier predicted a particular emotion given the actual emotion. The results demonstrate that the model was successful in classifying emotions, with high numbers of correct predictions for each emotion. Although there were some instances of misclassification, they were relatively infrequent. For instance, the model misclassified female sad as female disgust and male sad as male disgust. However, these misclassifications were not frequent, and the overall performance of the model was highly accurate.

Fig. 4.5 and Fig. 4.6 shows the classification report based on the performance of TCN with different feature combinations. Analysis of the precision, recall, and f1-score values for each class shows that the model performed exceptionally well across all classes. The weighted average F1-score, which is the overall F1-score taking into account class imbalance, is consistently high for all feature sets, ranging from 0.94 to 1.00. The precision values ranged from 0.93 to 1.00, except for the spectral contrast feature indicating a low false positive rate. Similarly, the recall values ranged from 0.89 to 1.00, indicating a low false negative rate. Additionally, the f1-score values ranged from 0.86 to 1.00, representing a harmonic mean between precision and recall.

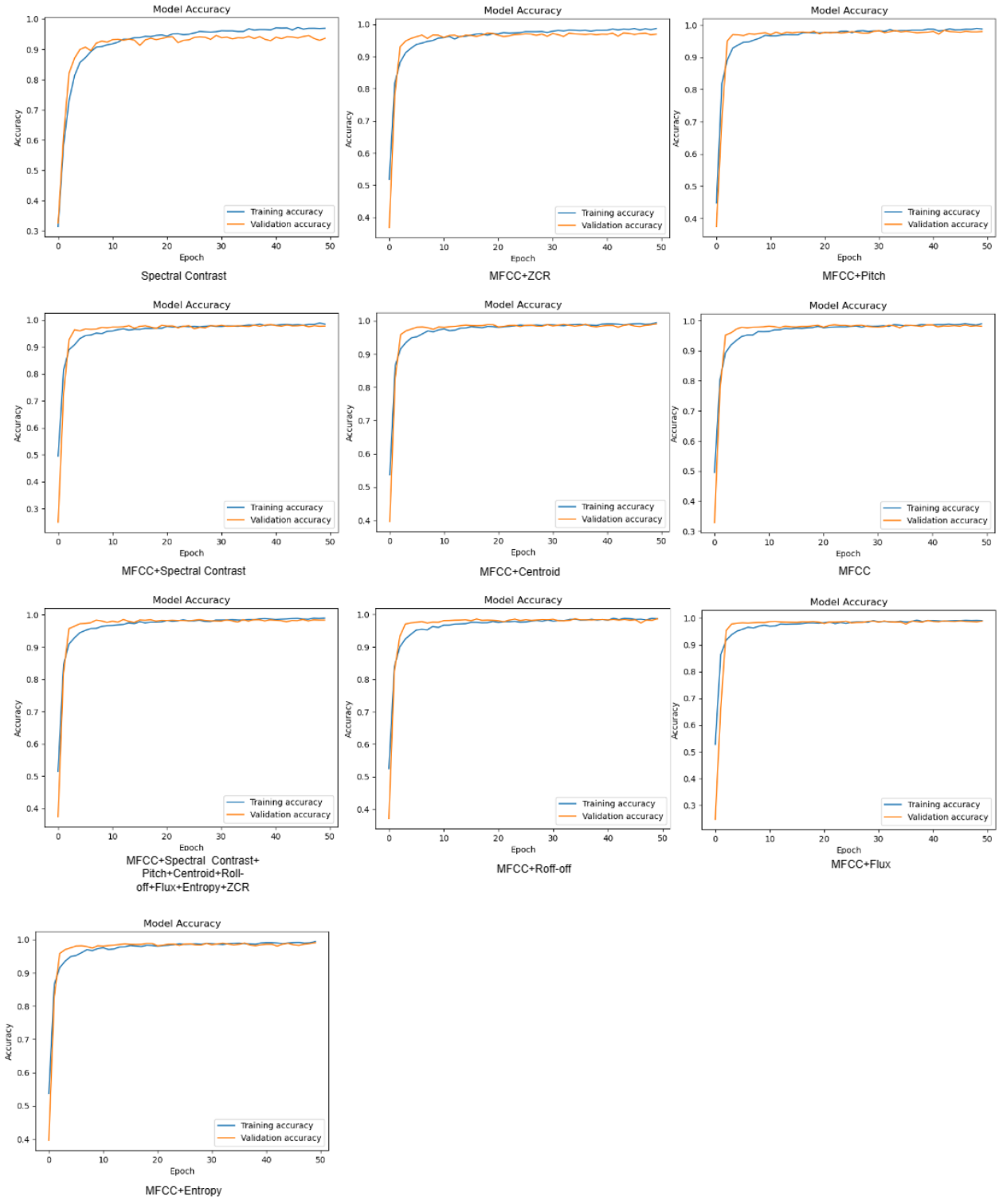

Figure 4.1: Accuracy plots based on the performance of TCN with different feature combinations

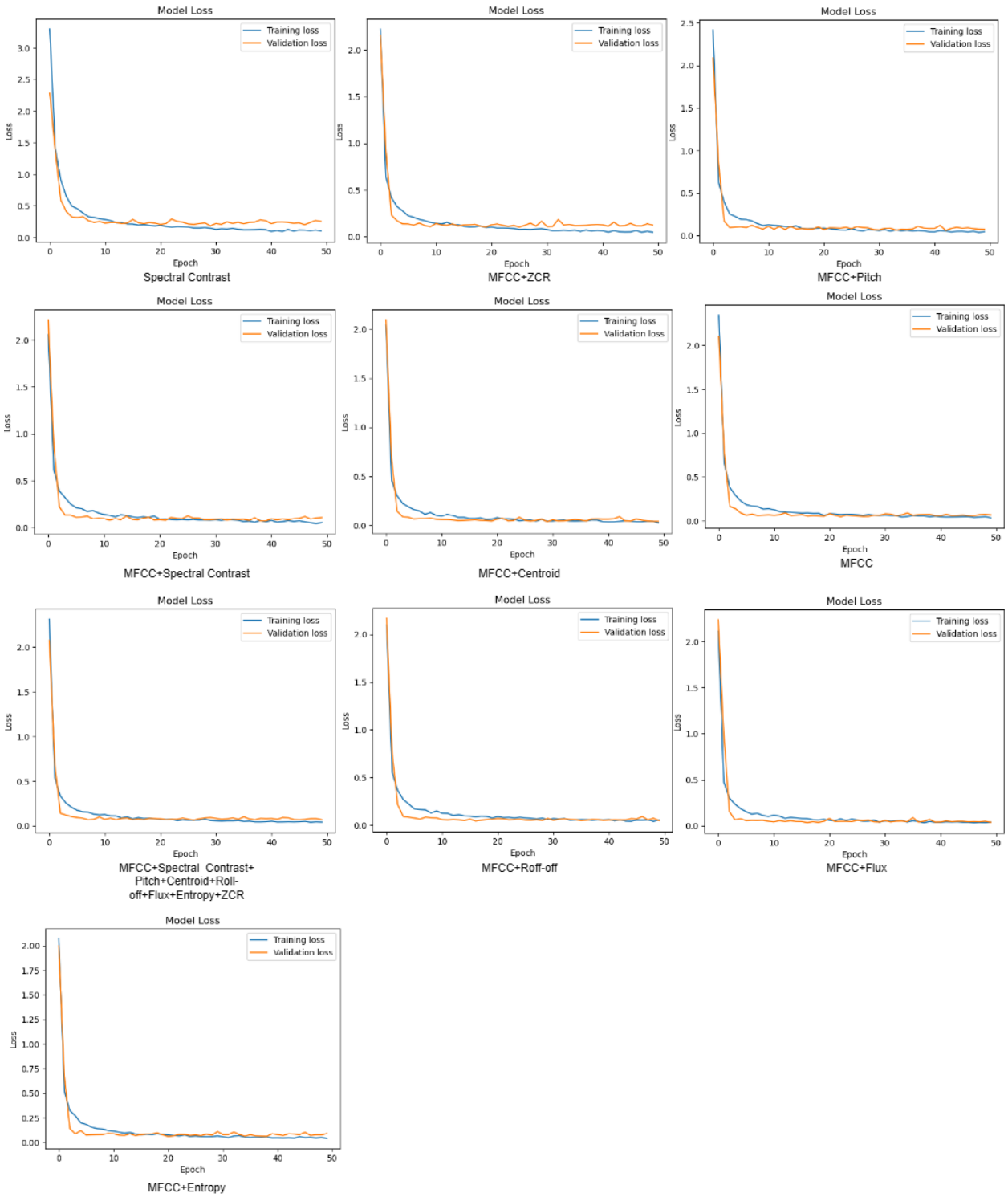

Figure 4.2: Loss plots based on the performance of TCN with different feature combinations

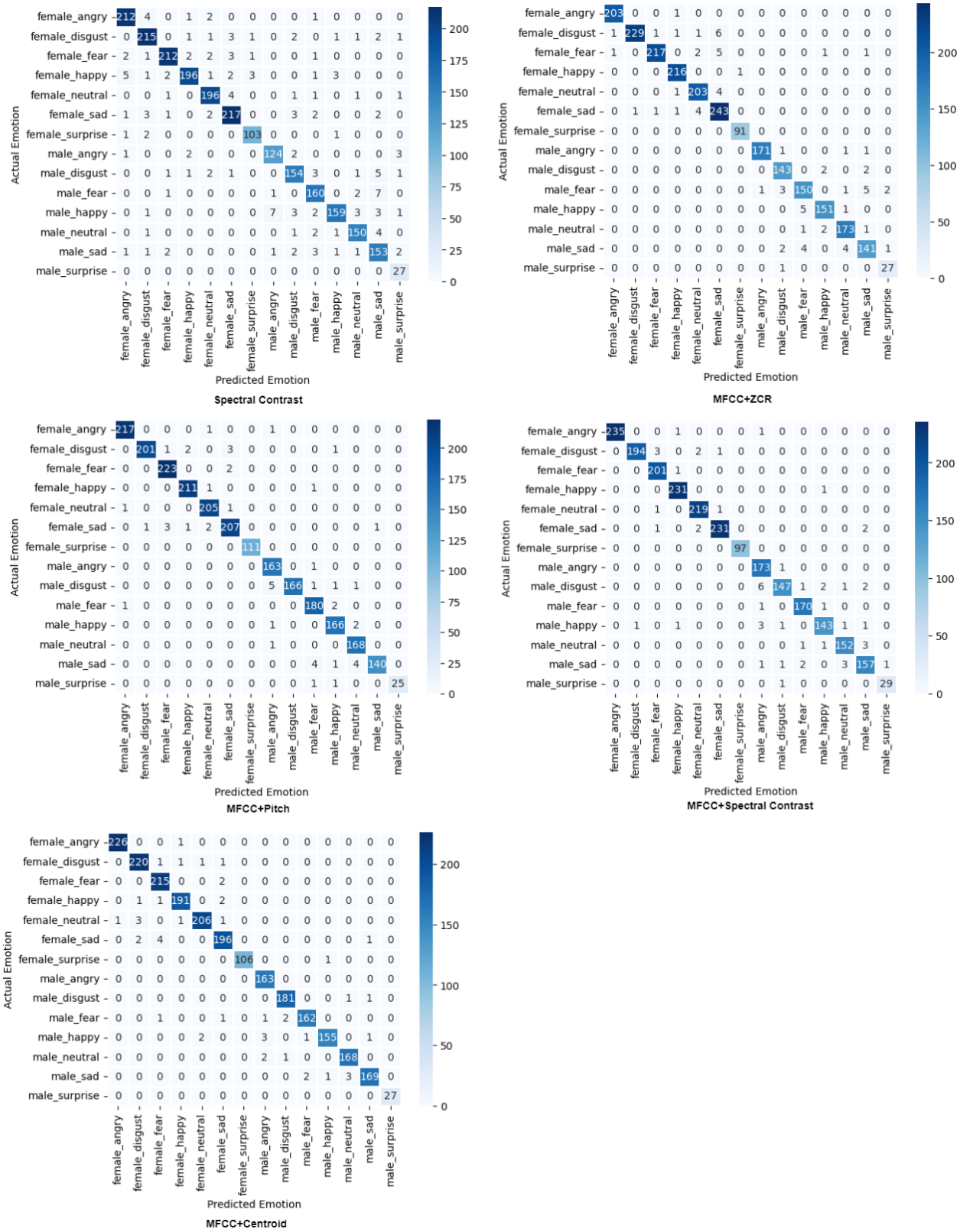

Figure 4.3: (a) Confusion matrix based on the performance of TCN with different feature combinations

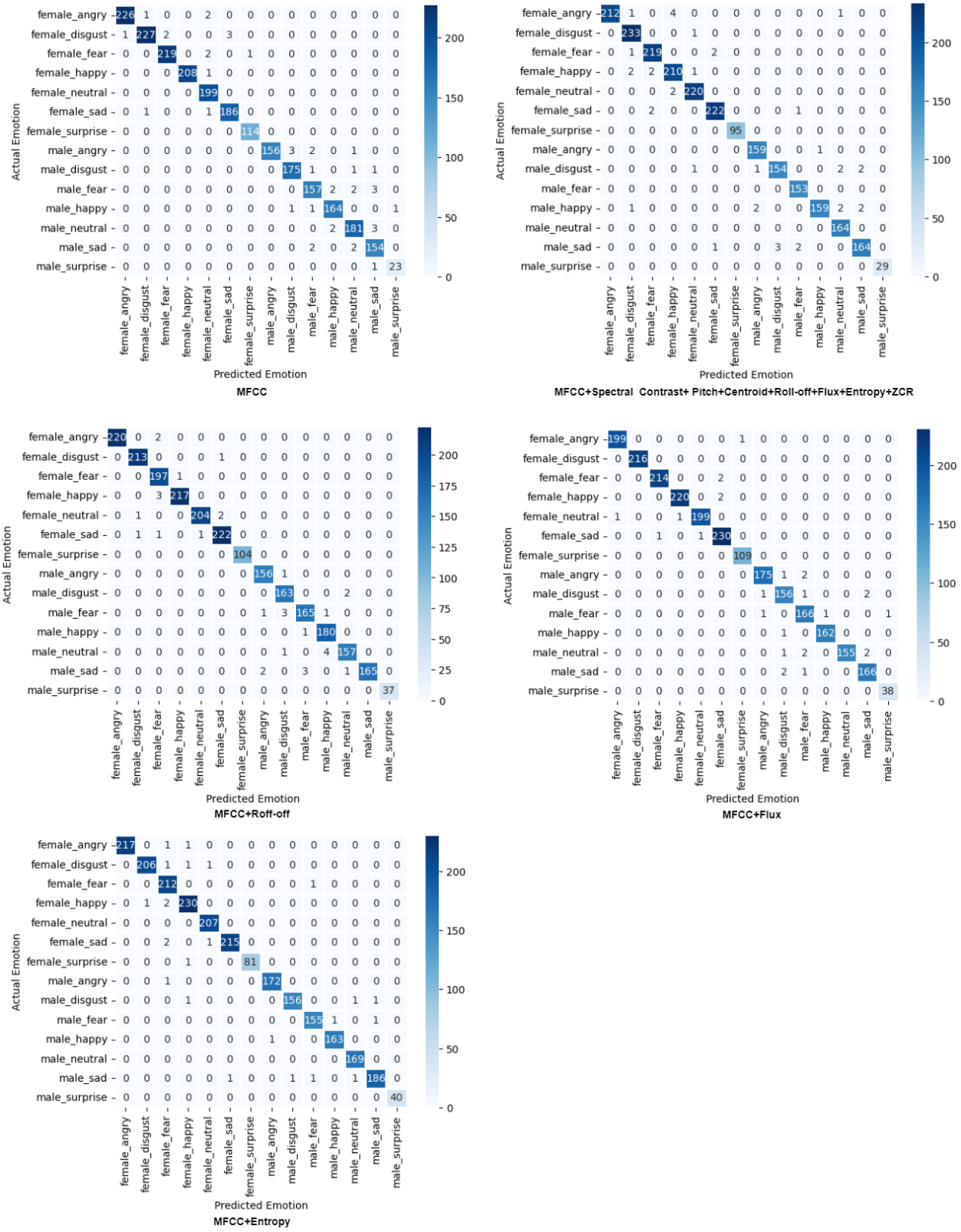

Figure 4.4: (b) Confusion matrix based on the performance of TCN with different feature combinations

	precision	recall	f1-score	support
female_angry	0.95	0.96	0.96	220
female_disgust	0.94	0.94	0.94	228
female_fear	0.96	0.95	0.95	224
female_happy	0.97	0.92	0.94	214
female_neutral	0.95	0.96	0.95	205
female_sad	0.94	0.94	0.94	231
female_surprise	0.95	0.96	0.96	107
male_angry	0.93	0.94	0.94	132
male_disgust	0.92	0.91	0.91	169
male_fear	0.91	0.94	0.92	171
male_happy	0.96	0.89	0.92	179
male_neutral	0.94	0.94	0.94	159
male_sad	0.87	0.92	0.89	167
male_surprise	0.75	1.00	0.86	27
accuracy			0.94	2433
macro avg	0.92	0.94	0.93	2433
weighted avg	0.94	0.94	0.94	2433

Spectral Contrast

	precision	recall	f1-score	support
female_angry	0.99	1.00	0.99	204
female_disgust	1.00	0.96	0.98	239
female_fear	0.99	0.96	0.97	227
female_happy	0.98	1.00	0.99	217
female_neutral	0.97	0.98	0.97	208
female_sad	0.94	0.97	0.96	250
female_surprise	0.99	1.00	0.99	91
male_angry	0.99	0.98	0.99	174
male_disgust	0.95	0.97	0.96	147
male_fear	0.94	0.93	0.93	162
male_happy	0.97	0.96	0.96	157
male_neutral	0.96	0.98	0.97	177
male_sad	0.93	0.93	0.93	152
male_surprise	0.90	0.96	0.93	28
accuracy			0.97	2433
macro avg	0.96	0.97	0.97	2433
weighted avg	0.97	0.97	0.97	2433

MFCC+ZCR

	precision	recall	f1-score	support
female_angry	0.99	1.00	0.99	219
female_disgust	1.00	0.97	0.98	208
female_fear	0.98	0.99	0.99	225
female_happy	0.99	0.99	0.99	213
female_neutral	0.98	0.99	0.99	207
female_sad	0.97	0.96	0.97	215
female_surprise	1.00	1.00	1.00	111
male_angry	0.95	0.99	0.97	164
male_disgust	1.00	0.95	0.98	174
male_fear	0.96	0.98	0.97	183
male_happy	0.97	0.98	0.97	169
male_neutral	0.96	0.99	0.98	169
male_sad	0.99	0.94	0.97	149
male_surprise	1.00	0.93	0.96	27
accuracy			0.98	2433
macro avg	0.98	0.98	0.98	2433
weighted avg	0.98	0.98	0.98	2433

MFCC+Pitch

	precision	recall	f1-score	support
female_angry	1.00	0.99	1.00	237
female_disgust	0.99	0.97	0.98	200
female_fear	0.98	1.00	0.99	202
female_happy	0.99	1.00	0.99	232
female_neutral	0.98	0.99	0.99	221
female_sad	0.99	0.98	0.99	236
female_surprise	1.00	1.00	1.00	97
male_angry	0.94	0.99	0.96	174
male_disgust	0.97	0.92	0.95	159
male_fear	0.98	0.99	0.98	172
male_happy	0.97	0.95	0.96	151
male_neutral	0.97	0.97	0.97	157
male_sad	0.95	0.95	0.95	165
male_surprise	0.97	0.97	0.97	30
accuracy			0.98	2433
macro avg	0.98	0.98	0.98	2433
weighted avg	0.98	0.98	0.98	2433

MFCC+Spectral Contrast

	precision	recall	f1-score	support
female_angry	1.00	0.99	1.00	227
female_disgust	0.97	0.98	0.98	224
female_fear	0.97	0.99	0.98	217
female_happy	0.98	0.98	0.98	195
female_neutral	0.99	0.97	0.98	212
female_sad	0.97	0.97	0.97	203
female_surprise	1.00	0.99	1.00	107
male_angry	0.96	1.00	0.98	163
male_disgust	0.98	0.99	0.99	183
male_fear	0.98	0.97	0.98	167
male_happy	0.99	0.96	0.97	162
male_neutral	0.98	0.98	0.98	171
male_sad	0.98	0.97	0.97	175
male_surprise	1.00	1.00	1.00	27
accuracy			0.98	2433
macro avg	0.98	0.98	0.98	2433
weighted avg	0.98	0.98	0.98	2433

MFCC+Centroid

Figure 4.5: (a) Classification report based on the performance TCN with different feature combinations

	precision	recall	f1-score	support
female_angry	1.00	0.99	0.99	229
female_disgust	0.99	0.97	0.98	233
female_fear	0.99	0.99	0.99	222
female_happy	1.00	1.00	1.00	209
female_neutral	0.97	1.00	0.99	199
female_sad	0.98	0.99	0.99	188
female_surprise	0.99	1.00	1.00	114
male_angry	1.00	0.96	0.98	162
male_disgust	0.98	0.98	0.98	178
male_fear	0.96	0.96	0.96	164
male_happy	0.98	0.98	0.98	167
male_neutral	0.97	0.97	0.97	186
male_sad	0.95	0.97	0.96	158
male_surprise	0.96	0.96	0.96	24
accuracy			0.98	2433
macro avg	0.98	0.98	0.98	2433
weighted avg	0.98	0.98	0.98	2433

MFCC

	precision	recall	f1-score	support
female_angry	1.00	0.97	0.99	218
female_disgust	0.98	1.00	0.99	234
female_fear	0.98	0.99	0.98	222
female_happy	0.97	0.98	0.97	215
female_neutral	0.99	0.99	0.99	222
female_sad	0.99	0.99	0.99	225
female_surprise	1.00	1.00	1.00	95
male_angry	0.98	0.99	0.99	160
male_disgust	0.98	0.96	0.97	160
male_fear	0.98	1.00	0.99	153
male_happy	0.99	0.96	0.98	166
male_neutral	0.97	1.00	0.98	164
male_sad	0.98	0.96	0.97	170
male_surprise	1.00	1.00	1.00	29
accuracy			0.98	2433
macro avg	0.98	0.98	0.98	2433
weighted avg	0.98	0.98	0.98	2433

MFCC+Spectral Contrast+ Pitch+Centroid+Roll-off+Flux+Entropy+ZCR

	precision	recall	f1-score	support
female_angry	0.98	1.00	0.99	216
female_disgust	0.96	0.98	0.97	219
female_fear	0.98	0.97	0.98	236
female_happy	0.99	0.99	0.99	207
female_neutral	1.00	0.99	1.00	206
female_sad	0.98	0.97	0.98	239
female_surprise	1.00	0.99	0.99	92
male_angry	0.99	0.96	0.97	153
male_disgust	0.94	0.98	0.96	163
male_fear	0.97	0.96	0.97	161
male_happy	0.99	0.95	0.97	160
male_neutral	0.96	0.97	0.97	185
male_sad	0.94	0.96	0.95	160
male_surprise	1.00	1.00	1.00	36
accuracy			0.98	2433
macro avg	0.98	0.98	0.98	2433
weighted avg	0.98	0.98	0.98	2433

MFCC+Roff-off

	precision	recall	f1-score	support
female_angry	0.99	0.99	0.99	200
female_disgust	1.00	1.00	1.00	216
female_fear	1.00	0.99	0.99	216
female_happy	1.00	0.99	0.99	222
female_neutral	0.99	0.99	0.99	201
female_sad	0.98	0.99	0.99	232
female_surprise	0.99	1.00	1.00	109
male_angry	0.99	0.98	0.99	178
male_disgust	0.97	0.97	0.97	160
male_fear	0.97	0.98	0.97	169
male_happy	0.99	0.99	0.99	163
male_neutral	1.00	0.97	0.98	160
male_sad	0.98	0.98	0.98	169
male_surprise	0.97	1.00	0.99	38
accuracy			0.99	2433
macro avg	0.99	0.99	0.99	2433
weighted avg	0.99	0.99	0.99	2433

MFCC+Flux

	precision	recall	f1-score	support
female_angry	1.00	0.99	1.00	219
female_disgust	1.00	0.99	0.99	209
female_fear	0.97	1.00	0.98	213
female_happy	0.98	0.99	0.99	233
female_neutral	0.99	1.00	1.00	207
female_sad	1.00	0.99	0.99	218
female_surprise	1.00	0.99	0.99	82
male_angry	0.99	0.99	0.99	173
male_disgust	0.99	0.98	0.99	159
male_fear	0.99	0.99	0.99	157
male_happy	0.99	0.99	0.99	164
male_neutral	0.99	1.00	0.99	169
male_sad	0.99	0.98	0.98	190
male_surprise	1.00	1.00	1.00	40
accuracy			0.99	2433
macro avg	0.99	0.99	0.99	2433
weighted avg	0.99	0.99	0.99	2433

MFCC+Entropy

Figure 4.6: (b) Classification report based on the performance TCN with different feature combinations

Chapter 5

Conclusion

The goal of the work was to classify emotions from audio files into 14 classes based on gender. To achieve this, a combination of different features and a Temporal Convolutional Network (TCN) classifier were utilized. Features extracted, include MFCCs, spectral contrast features, pitch, spectral centroid, spectral roll-off, onset flux, spectral flatness entropy, and zero crossing rate (ZCR). However, these features did not capture the entire temporal structure of the audio signal. Thus, TCN was incorporated to capture long-term dependencies in the audio signal. The work showed that the combination of MFCC and entropy achieved the highest training and validation accuracies, with 99.4% and 99.01%, respectively. The other combinations of features also achieved high accuracy rates, indicating the effectiveness of using TCN in conjunction with a variety of feature sets to classify emotions. Overall, the incorporation of TCNs into the feature extraction process enhances the performance of the emotion classification model. The TCN model can learn additional features that capture long-term dependencies in the audio signal, which are not captured by the spectral features alone. This provides a more complete representation of the audio signal and improves the accuracy of the emotion classification model. Future studies could investigate the effectiveness of the proposed approach with larger datasets and explore the potential of using other deep learning models in emotion classification from audio signals.

References

[1] 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*, vol. 7, pp. 117 327–117 345, 2019.

[2] A. Koduru, H. B. Valiveti, and A. K. Budati, "Feature extraction algorithms to improve the speech emotion recognition rate," *International Journal of Speech Technology*, vol. 23, no. 1, pp. 45–55, 2020.

[3] M. Chen, X. He, J. Yang, and H. Zhang, "3-d convolutional recurrent neural networks with attention model for speech emotion recognition," *IEEE Signal Processing Letters*, vol. 25, no. 10, pp. 1440–1444, 2018.

[4] D. Issa, M. F. Demirci, and A. Yazici, "Speech emotion recognition with deep convolutional neural networks," *Biomedical Signal Processing and Control*, vol. 59, p. 101894, 2020.

[5] P. Jiang, H. Fu, and H. Tao, "Speech emotion recognition using deep convolutional neural network and simple recurrent unit." *Engineering Letters*, vol. 27, no. 4, 2019.

[6] M. Farooq, F. Hussain, N. K. Baloch, F. R. Raja, H. Yu, and Y. B. Zikria, "Impact of feature selection algorithm on speech emotion recognition using deep convolutional neural network," *Sensors*, vol. 20, no. 21, p. 6008, 2020.

[7] A. Aggarwal, A. Srivastava, A. Agarwal, N. Chahal, D. Singh, A. A. Alnuaim, A. Alhadlaq, and H.-N. Lee, "Two-way feature extraction for speech emotion recognition using deep learning," *Sensors*, vol. 22, no. 6, p. 2378, 2022.

[8] M. B. Er, "A novel approach for classification of speech emotions based on deep and acoustic features," *IEEE Access*, vol. 8, pp. 221 640–221 653, 2020.

[9] M. Sajjad, S. Kwon *et al.*, "Clustering-based speech emotion recognition by incorporating learned features and deep bilstm," *IEEE access*, vol. 8, pp. 79 861–79 875, 2020.

[10] S. R. Livingstone and F. A. Russo, "The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english," *PloS one*, vol. 13, no. 5, p. e0196391, 2018.

[11] P. Jackson and S. Haq, "Surrey audio-visual expressed emotion (savee) database," *University of Surrey: Guildford, UK*, 2014.

[12] M. K. Pichora-Fuller and K. Dupuis, "Toronto emotional speech set (tess)," *Scholars Portal Dataverse*, vol. 1, p. 2020, 2020.

[13] 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*, vol. 5, no. 4, pp. 377–390, 2014.

[14] P. P. Singh and P. Rani, "An approach to extract feature using mfcc," *IOSR Journal of Engineering*, vol. 4, no. 8, pp. 21–25, 2014.

[15] S. Kumar and S. Thiruvenkadam, "An analysis of the impact of spectral contrast feature in speech emotion recognition." *Int. J. Recent Contributions Eng. Sci. IT*, vol. 9, no. 2, pp. 87–95, 2021.

[16] 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*. American Society for Engineering Education, 2008, pp. 1–7.