Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 258 (2025) 3693–3702

Procedia Computer Science

www.elsevier.com/locate/procedia

International Conference on Machine Learning and Data Engineering

Speech Emotion Recognition using Mel Spectrogram and Convolutional Neural Networks (CNN)

Vidhi Sareen[1], Seeja K.R.

Department of Computer Science and Engineering, Indira Gandhi Delhi Technical University for Women, India

Abstract

Humans rely on emotions to express thoughts and engage in everyday activities such as interaction, learning, and decision-making. Speech Emotion Recognition (SER) systems process audio inputs to detect emotions, with applications in sectors like customer service, healthcare, education, and market research. This paper presents a method for SER using Mel spectrograms and Convolutional Neural Networks (CNNs). Audio samples are converted into Mel spectrogram images representing the frequency spectrum on the Mel scale and these images are then used to train a CNN model. The system predicts emotions such as anger, fear, disgust, neutral, happy, surprise, and sad. We evaluate the model using RAVDESS, SAVEE, TESS, and the combined dataset, attaining accuracies of 70% for RAVDESS, 60% for SAVEE, 99.89% for TESS, and 87% for the combined dataset. This approach demonstrates the effectiveness of combining Mel spectrograms with CNNs for robust emotion recognition from speech.

Keywords: Speech Emotion Recognition (SER), Mel Spectrogram, Convolution Neural Networks (CNN)

1. Introduction

Emotions can be expressed in different forms like text, speech, facial expressions, or body posture. Speech is an elementary mode of communication for humans through which they share their emotions. The process of detecting or recognizing emotions through speech or audio signals is called speech emotion recognition (SER). Emotion recognition using speech signals has become increasingly important in different areas like lie detection, security

[1] Vidhi Sareen Tel.: +91-981-881-3600
E-mail address: vidhi022mtcse23@igdtuw.ac.in

1877-0509
10.1016/j.procs.2025.04.624

concerns, emergency call centres, e-learning, smart classrooms [1], and health care analysis [2]. The success of the Speech Emotion Recognition (SER) system rely on the features they are extracting and how are they classified. Choosing the best set of features is difficult because different features depict different emotions.

The motivation behind this study is to improve the detection of emotions by leveraging not just raw audio but also spectral representations. The effective SER system integrates innovative features, extraction techniques, and advanced Learning techniques. For example, features like 'Mel-Frequency Cepstral Coefficient (MFCC)', 'Chromagram', 'Tonnetz Representation', and 'Spectral Contrast' features are derived from raw audio signals and passed to Convolutional Neural Networks (CNNs) for classification demonstrating a marvelous improvement of the SER system [3]. In paper [4] it explains how spectral and Prosodic features like pitch, loudness, and frequency along with machine learning models like SVM, RBG, and Back Propagation were used to detect emotions from speech. The increasing interest in Speech emotion recognition (SER) is due to advancements in Machine Learning and Deep Learning. It has demonstrated outstanding success in handling complex tasks.

Extraction of features is a critical component in SER because it directly influences the model performers. The most effective features used by the researchers so far include MFCCs, Mel spectrogram, pitch, energy, ZCR etc. Recent advancements have introduced more sophisticated methods such as the IS10 feature set [5], which incorporates spectral and prosodic features, and the Multitaper Mel Frequency Spectrogram (MTMFS) [6], which enhances frequency resolution and stability in noisy conditions. Many features like MFCCs and Mel spectrogram are commonly used for speech processing and classification.

The evolution of machine learning algorithms has also extensively impacted SER. Early techniques like SVM, HMM, and Decision Trees have been used for emotion classification tasks. With the progress in Deep Learning, there have been advancements in CNNs [3], LSTM networks [7], and Transformer [18], [20] architectures have been used which have brought substantial improvements in accuracy and robustness.

Data augmentation techniques have been crucial in advancing SER systems. Approaches like tempo and speed perturbation [8] tackle issues of limited and variable data, enabling models to perform more robustly under diverse speech conditions. Additionally, ensemble models, which aggregate predictions from several classifiers, enhance emotion classification accuracy by combining multiple viewpoints and mitigating overfitting. Although progress has been made in the field of SER, it still faces several challenges. Issues like variability in speech data, linguistic differences in emotional expression, and many more continue the research journey in SER.

In this study, the Mel spectrogram and Convolution Neural Network are used to detect emotions from a speech. We have converted the audio samples into Mel Spectrogram Images to get a better understanding of the audio. We also have used data augmentation techniques and CNN model to detect the emotions of the audio. Our model is evaluated on different datasets.

2. Related Work

We have reviewed previous research done in the field of Speech Emotion Recognition (SER). Issa et al.[3] have used different datasets where they have worked with the raw sound data. They have audio features like 'MFCC', 'Chromagram', 'Mel-Scale Spectrogram', 'Tonnetz', and 'Spectral contrast' features. The methodology involved extracting five different spectral features from sound files, stacking the resulting matrices into a one-dimensional array (by calculating mean), and feeding this array into a 1-D CNN. Zhao et al. [13] have also used 1-D CNN along with LSTM to detect emotions from audio clips.

Extracting features is a crucial aspect that can improve the speech emotion recognition rate as discussed by Koduru et al. [14]. They have used the Global feature algorithm to remove redundant information from the feature and identify emotions from it. Many other techniques are used for extracting statistical features [15] by using a broader range of Standard deviation (SD) values. Identifying which features will impact which emotion is a difficult task. Many audio features are being used in the SER like pitch, energy, Zero Crossing Rate (ZCR), MFCC, Chromagram, Spectrogram, etc. Ancilin et al. [16] have proposed 2 modifications in the extraction of MFCC like it has used energy spectrum instead of spectrum and it doesn't take DCT into account and extracted Mel Frequency Magnitude Coefficient.

Many researchers have used traditional machine learning techniques like SVM [15], decision trees, etc. Aouani et.al. [15] have extracted a 42-dimensional features vector of the audio features and then SVM for classification. With the advancement in technology, many Deep Learning methods have also been applied in the field of SER. Zhang et al [5] have explored an autoencoder with an emotion embedding method to extract deep emotion features whereas

some researchers have also used 2D-CNN and eXtreme Gradient-boosting techniques to detect emotions [9].

Praseetha et al. [8] have proposed a model with Gated Recurrent Unit Networks (GRU) and filterbank energies as input. Chen et al. [17] have introduced a BiLSTM network along with an attention mechanism (dual Attention-BLSTM). In this MFCC features (including deltas and deltas-deltas) are used and Log Mel-spectrograms are extracted from CNN and are sent into the BLSTM network. Many hybrid models for SER, like LSTM and Transformer, are used for understating long-term audio patterns and identifying emotions. MFCC is extracted from the speech audio as input features [18].

Pham et al. [19] have given an approach combining attention-based convolutional-recurrent neural networks and a data augmentation method for Speech Emotion Recognition. They have used traditional and Generative Adversarial Networks (GAN) for data augmentation. Through the application of 'Convolutional-Recurrent Neural Network (CRNN), they have extracted high-level representations from Mel Spectrogram features. Singh et al. [21] have proposed Gender Dependent training model for detecting emotions. It has used only MFCC features and its variants (delta MFCC and delta-delta MFCC) which are compared with the baseline model. It has shown that by taking gender attributes the accuracy has been improved by 6.90%.

Yao et al. [22] It has proposed a fusion of three multi-tasking learning-based classifiers. It has used three classifiers. The first one is High-level Statistical Functions (HSFs) fed into a DNN, the second one is Spectrograms processed by a CNN and the third one is Low-level descriptors (LLDs) processed by an RNN. The results from these classifiers are combined at the decision level using confidence scores generated by each model. Table 1 shows a comparative study of the research done in the field of SER.

Table 1. Comparison of different existing SER research done

Author	Dataset Used	Preprocessing	Audio Parameters / Features	Model / Approach	Accuracy (%)
[3]	RAVDESS, EMO-DB, IEMOCAP	Work directly with raw audio data	MFCC, Chromagram, Mel-scale spectrogram, Tonnetz, spectral contrast features	1-D CNN	RAVDESS – 71.61 EMO-DB – IEMOCAP – 64.30
[5]	IEMOCAP, EMODB	log magnitude spectrogram of the audio signal are used	Log magnitude spectrogram and IS10 feature set from openSMILE toolkit	Autoencoder with emotion embedding	IEMOCAP – 71.2 EMODB – 95.6
[8]	TESS	audio data undergoes data augmentation techniques including tempo perturbation and speed perturbation	primary feature used is Filterbank Energies	Gated Recurrent Unit (GRU)	average accuracy is 93
[9]	RADVESS	Pre-processing involves trimming, zero padding. Segmentation	MFCC	2D-CNN + XGBoost	96.47
[14]	RAVDESS	Audio are pre-processed by removing noise using filters	MFCC, DWT, Pitch, Energy, ZCR	SVM, Decision Tree, LDA	SVM – 70, Decision Tree – 85, LDA – 65
[15]	RML	Used audio signals	MFCC, HNR, ZCR, TEO	Autoencoder + SVM	74.07
[16]	EMO-DB, RAVDESS, SAVEE, EMOVO, eNTERFACE , Urdu databases	Speech samples are framed, Fourier-transformed, and converted to Mel scale and further filtering and logarithmic compression are used	MFMC	SVM	EMO-DB – 81.50, RAVDESS – 64.31, SAVEE – 75.63, EMOVO -73.30, eNTERFACE – 56.41, Urdu databases – 95.25

In Table 1, a detailed comparison of preprocessing techniques, audio features, models, and approaches used by various researchers is presented along with their accuracies. While these techniques have shown significant improvement in speech emotion recognition systems but there are still some limitations. One of the main limitations is that most of the research relies on raw audio signals which can introduce noise or variability that can affect emotion recognition. To address this, we have opted for an approach in which audio signals are converted into Mel spectrogram images. This approach leverages the visual representation of audio data, allowing for more effective feature extraction and improved classification of emotions.

3. Methodology

The methodology used in this research is shown in Fig 1.

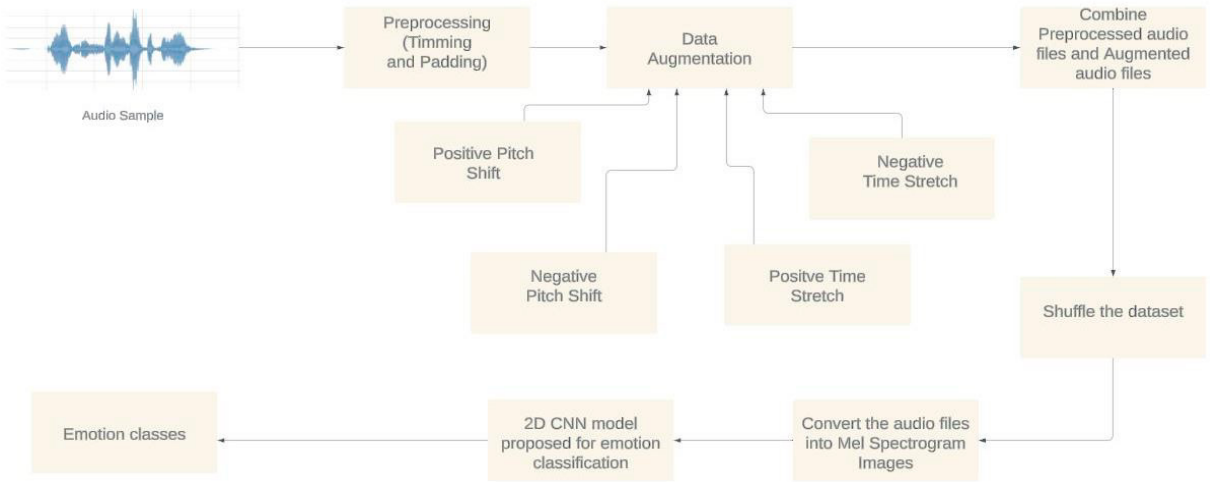

Fig 1. Block diagram of the proposed methodology

First, the audio samples were preprocessed by trimming the audio files to a standard-length of 2500 milliseconds to ensure that all the files are the same length. This trimming ensures consistency across the dataset while capturing the relevant portion of the audio that contains emotional information. Next, data augmentation techniques are used on the preprocessed data to expand the dataset and introduce variability. Four data augmentation techniques used are Positive pitch shift, which increases the pitch, Negative pitch shift, which decreases the pitch without the change in duration, Positive time stretch which expands the duration and Negative time stretch which compresses the duration. These augmentation techniques enhance the diversity of the dataset. After augmentation, the preprocessed audio files and augmented files are combined to create a large dataset. The combined dataset is then shuffled to ensure that all the emotion labels are evenly distributed, reducing bias during training. Then these audio files are converted into Mel Spectrogram images, which are time-frequency representations of the sound. These spectrogram images serve as an input to the proposed 2D CNN model for predicting the emotions. The proposed 2D CNN model is shown in Fig 2.

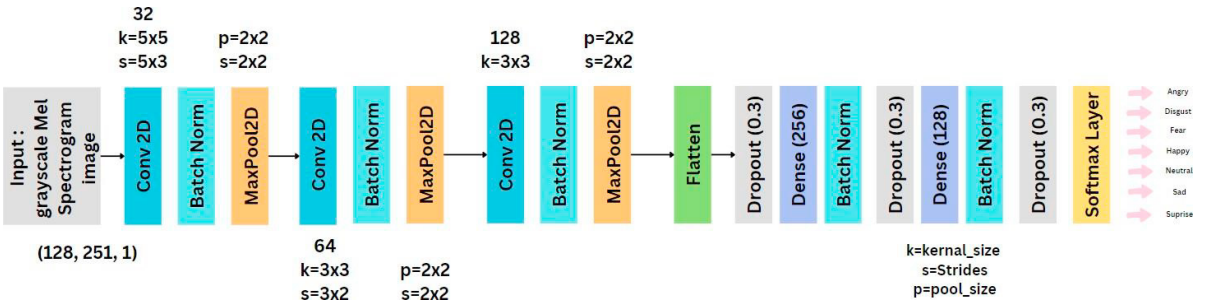

Fig 2. Proposed CNN architecture

4. Implementation

4.1. *Datasets*

We have used four datasets in our research. Those are RAVDESS, SAVEE, TESS, and a Combined Dataset. Detailed information about each dataset is provided below.

The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) was chosen because it is easily available [23]. The dataset has audio and video recordings. It consists of twenty-four actor's voice samples where 12 are male and 12 are female. They are narrating two English sentences with eight different emotions which is efficient for emotion analysis [10]. For our project, we have chosen 7 emotions which are Happy, Angry, Sad, Disgust, Neutral, Fear, and Surprise. There are a total of 1440 recordings.

The second dataset we chose is the Surrey Audio-Visual Expressed Emotion (SAVEE) dataset. It includes audio and visual data, but we only used the audio part. The SAVEE dataset has recordings of seven emotions which are neutral, Angry, Surprise, Happy, Fear, Sad, and Disgust. Four male actors recorded 480 sentences in British English, with each sentence reflecting one of these emotions. The Toronto Emotional Speech Set (TESS) dataset was chosen as the third dataset. The Tess dataset has recordings of seven emotions which are Anger, Disgust, Fear, Happy, Surprise, Sad, and Neutral. It features voices from two women, aged 26 and 64. The dataset has 2800 audio files. However, using only TESS to train a speech emotion recognition (SER) model isn't ideal because it has recordings from only two people. Combining TESS with other datasets can solve this problem and make the model better at recognizing emotions from different voices [10].

By combining three benchmark datasets, we created a larger and more diverse multi-speaker dataset with distinct emotions. We did this to increase the data's size and variety, making it more robust and better at capturing different emotional expressions from various speakers, languages, and recording conditions. We have combined the dataset based on emotion labels. This method of combining different speech emotion datasets resulted in a well-structured dataset.

4.2. *Preprocessing and Data Augmentation*

In this research, we need to preprocess audio data to make sure it's consistent and of good quality before we analyze it. We have a dataset with audio file paths and their emotion labels stored in a CSV file. Using Python, we start by loading these audio files. Each file is trimmed to a standard length of 2500 milliseconds, starting from 500 milliseconds in, to keep everything uniform [9]. If any segment is shorter than this length, we add zero padding to make up the difference. The edited files are then saved in WAV format to a specific output folder. This process is done in batches, where each file's path is read from the CSV, and the processed files are saved with a new path. Finally, we check that both the original and edited audio lengths are consistent to ensure quality.

We have then performed data augmentation to expand our dataset. Many data augmentation techniques are used in Speech Emotion Recognition [13],[19]. The data is augmented either on the images or on the audio. We have performed four techniques in our research. They are Positive Pitch Shift (PPS), Negative Pitch Shift (NPS), Slow Time Stretches (STS), and Fast Time Stretches (FTS) on the audio samples [12]. Pitch shifting is a technique that alters the original pitch of a sound, raising or lowering it, while keeping its duration unchanged. Positive Pitch Shift

increases the pitch were as negative pitch shift decreases the pitch. In our research, we have done a positive pitch shift by the factor of +3 and a negative pitch shift by the factor of -3. For example, Fig. 3 shows waveform of the positive and negative pitch shift of a sample audio.

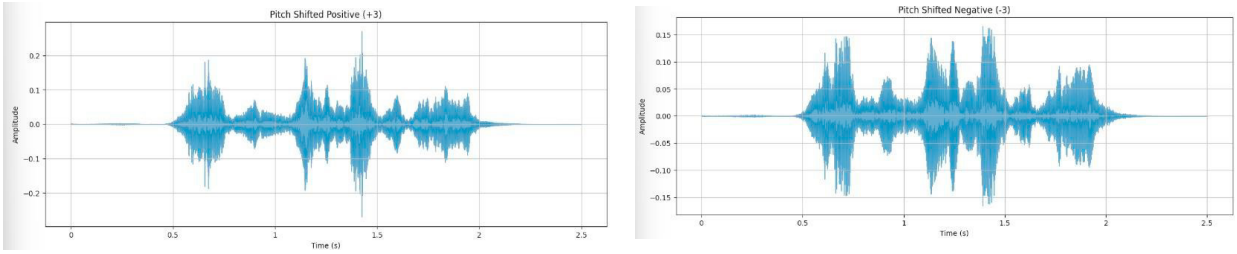

Fig. 3. (a) Positive Pitch Shift (PPS); (b) Negative Pitch Shift (NPS)

Time stretching is a technique that alters the speed or duration of a sound while preserving its pitch Slow time stretch, stretches the audio signal whereas fast time stretch compresses the duration. In our research, we have applied a 0.8 factor for Slow Time Stretch and 1.3 for Fast Time Stretch. For example, Fig. 4 shows waveform of the positive and negative pitch shift of a sample audio. For example, Fig. 4 shows the waveform of the slow and fast time stretch of a sample audio.

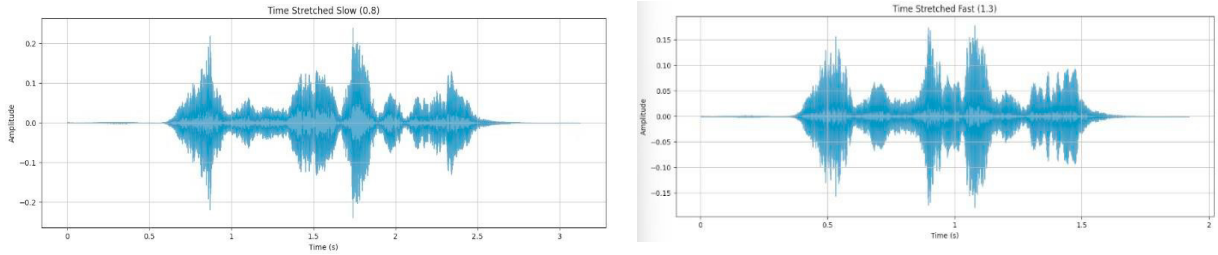

Fig. 4. (a) Slow Time Stretch (STS); (b) Fast Time Stretch (FTS)

4.3. Speech Feature for Deep Neural Network Training

The augmented audio files were converted into a Mel-spectrogram. A spectrogram is a tool that depicts how the signal's frequencies change over time. The Mel scale, designed to align with the human auditory system, converts Hertz to Mels using the following Eq. (1), where m represents Mels and f represents Hertz. Here, lower frequencies are represented with narrower intervals and higher frequencies with wider intervals.

$$m = 2595\,(1 + \frac{f}{700}) \tag{1}$$

We use Mel spectrograms to provide our models with sound information similar to what humans perceive. We have used the Librosa library to convert the audio signal to Mel spectrogram and then transform amplitude squared values to decibels to standardize the representation which enhances the spectrogram's visual representation. We have taken sample rate as 44.1kHz and 128 Mel bands, Fast-Fourier Transforms window size as 2048, and hop length as 512. Each resulting sample has a shape of 128 by 251 where 128 is the Mel bands and 251 is the Mel frequency bins or features in the spectrogram. We have performed normalization as it impacts the accuracy of Speech Emotion Recognition and can detect emotions very well [11]. Fig 5 shows the Mel Spectrogram of different emotions.

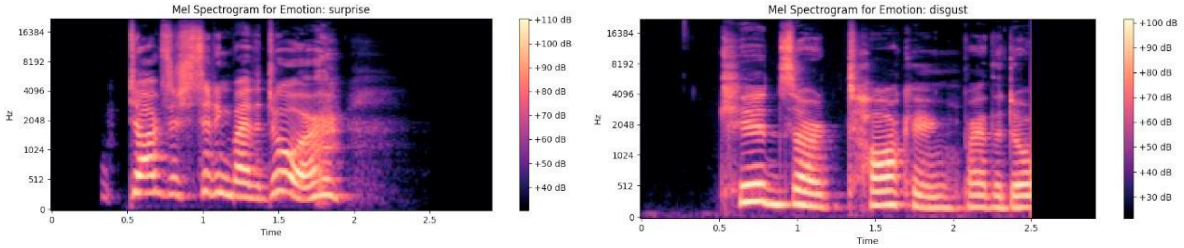

Fig. 5 (a) Mel Spectrogram for Surprise Emotion; (b) Mel Spectrogram for Disgust Emotion

4.4. Implementation of the proposed CNN Model

The proposed Convolutional Neural Network (CNN) model is used to detect emotions from speech and has layers that perform convolutions, batch normalization, max-pooling, and dropout. The model's input layer accepts data size as (128 × 251 × 1), which is a grayscale image. The first convolutional layer has 32 filters, each with a size of (5 × 5), and processes the image with a stride of (5 × 3). After applying this layer, batch normalization and the Rectified Linear Unit (ReLU) function are used to activate the output. We apply a max-pooling operation with a pool size of (2 × 2) to reduce the size of the feature maps. The second convolutional layer has 64 filters with a size of (3 × 3) and uses strides of (3 × 2). This layer also includes batch normalization and ReLU activation. Next is a third convolutional layer with 128 filters of size (3 × 3), followed by batch normalization and ReLU activation. We use 'same' padding in this layer to ensure the output feature map has the same dimensions as the input. The output then goes through another max-pooling layer with a pool size of (2 × 2). Afterward, the feature maps are flattened into a one-dimensional vector. The model includes dense layers with (256) and (128) units, with dropout to prevent overfitting. Each dense layer is activated by the ReLU function, with batch normalization and dropout at a rate of 0.3. The final layer is a dense layer with a Softmax activation function, which helps in classification. The model is compiled using the Adam optimizer, with a learning rate of 0.0001 and a batch size of 32. It uses sparse categorical cross-entropy as the loss function, suitable for multi-class classification, and tracks accuracy as the performance metric. To enhance training and reduce overfitting, we use an EarlyStopping callback.

5. Experimental Results

We have performed different experiments on the proposed model with the help of different datasets. We experimented on different datasets that contained audio files. The total data was split into an 80:20 ratio. Normalization and reshaping were performed before training the model over 70 epochs. We have used an early stopping technique to avoid overfitting. We have evaluated the accuracy and F1-score of different emotions.

For the combined dataset, the original audio files were 4720 and after data augmentation, the total number of audio files came to about 23600. These audio files were shuffled to ensure that each batch represents samples of the entire dataset. The audio files were converted into Mel spectrogram images and these images were processed along with their labels for input for the proposed model. The model got trained for 29 epochs and was terminated to avoid overfitting with an accuracy of about 87 %. Fig 6 (a) shows the graph between the accuracy and epoch and Fig 6 (b) shows the graph between loss and epoch. The model was able to recognize most of the emotions very well with an F1 score ranging between 0.82 to 0.88. The ‘Surprise’ emotion has the highest F1-score of about 0.90.

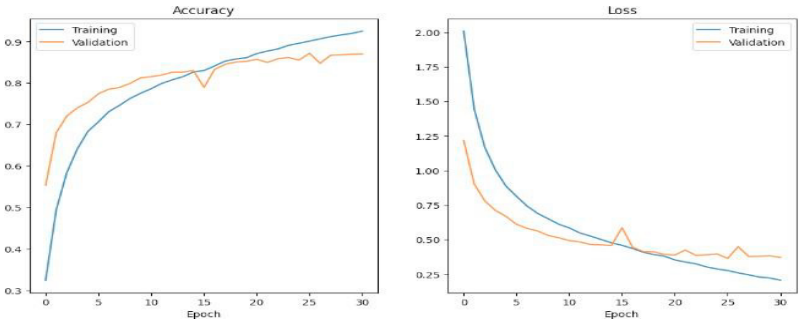

Fig. 6. (a) Accuracy vs Epoch for Combined Dataset; (b) Loss vs Epoch for Combined Dataset

For the RAVDESS dataset, the original audio files were about 1440 but after preprocessing and data augmentation the dataset was expanded to 7200 files. These audio files were converted into images and then processed as input for the proposed model. This model was trained to 70 epochs but observed that it got terminated at 46 epochs to avoid overfitting. The model accuracy comes to about 70% with varying Precision, Recall, and F1 scores across different emotions. Fig 7 (a) shows the graph between the accuracy and epoch and Fig 7 (b) shows the graph between loss and epoch highlighting its strength in recognizing 'Neutral' and 'Angry' emotions and its challenges with 'Happy' and 'Sad' emotions.

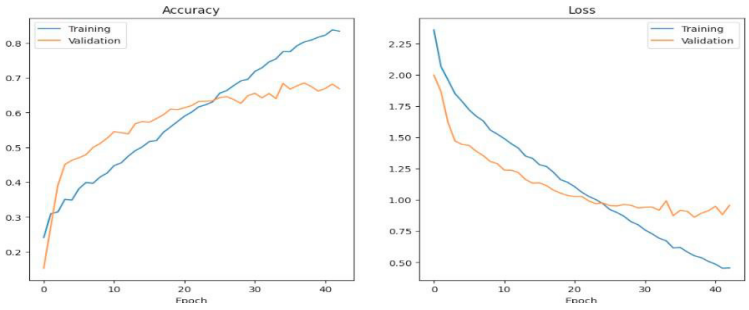

Fig. 7. (a) Accuracy vs Epoch for RAVDESS dataset; (b) Loss vs Epoch for RAVDESS Dataset

For the SAVEE dataset, the original audio files were about 480 and after data augmentation, the total audio files came to about 2400. We then converted these 2400 audio files into Mel Spectrogram images and processed them for input for the CNN model. To prevent overfitting the model got terminated at 42 epochs with the accuracy of 60%. Our model was able to recognize neutral emotion with an F1 score of 0.76, while ‘Disgust’ had the lowest at 0.47. Fig 8 (a) shows the graph between the accuracy and epoch and Fig 8 (b) shows the graph between loss and epoch

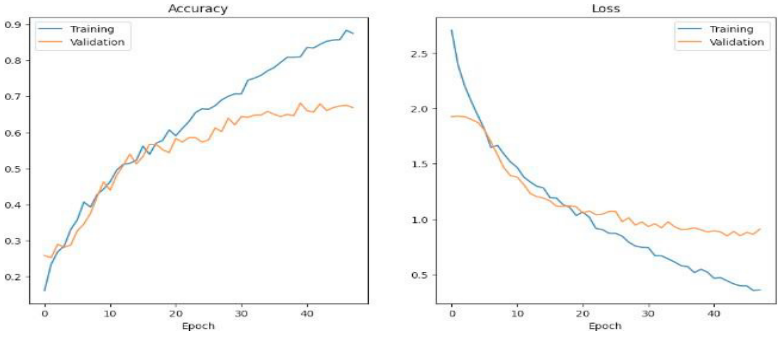

Fig. 8. (a) Accuracy vs Epoch for SAVEE Dataset; (b) Loss vs Epoch for SAVEE Dataset

For the TESS dataset, 2800 were the original audio files and after data augmentation, the total audio files came to about 14000. We have then converted these audio files to Mel Spectrogram and then transformed them into an array. These are then passed to the CNN model and achieved an accuracy of 99.89%. Fig. 9 (a) shows the graph between the accuracy and epoch and Fig. 9 (b) shows the graph between loss and epoch for the TESS dataset.

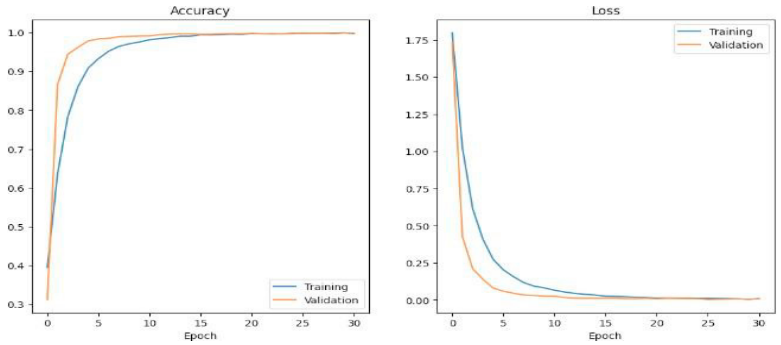

Fig. 9. (a) Accuracy vs Epoch for TESS Dataset; (b) Loss vs Epoch for TESS Dataset

While the result shows a strong performance of our CNN model but some limitations need to be considered. The variation in the dataset sizes and the diversity of the emotions may impact the model's generalizability. Certain emotions with low accuracy are the potential areas for improvement in future studies.

The proposed emotion detection model is compared with other state-of-the-art models and is shown in Table 2.

Table 2. Comparative table between our proposed emotion detection model and other systems.

State of the art	Dataset	Accuracy (%)
[3]	RAVDESS	71.61
[8]	TESS	93
[16]	RAVDESS, SAVEE	81.50, 75.63
[21]	RAVDESS	72.07
Our Proposed model	RAVDESS	70
	SAVEE	60
	TESS	99.89
	Combined dataset (RAVDESS + TESS + SAVEE)	87

6. Conclusion

This paper presents a method and design to identify emotions from speech. This method involves Mel Spectrograms and Convolutional Neural Networks. We have trained and tested our model on different datasets for Speech Emotion Recognition (SER). Our approach, which converts audio samples into Mel Spectrogram images for training, demonstrates the potential of CNNs in accurately classifying emotions such as anger, fear, disgust, neutral, happy, surprise, and sadness. The significance of this research lies in contributing to SER by providing an effective approach that is beneficial for various applications. While the results are promising, there are several areas for future research to address the limitations of the basic CNN model. For instance, basic CNN architectures may struggle with generalization due to limited model complexity. Future research could explore advanced architecture like Residual Networks (ResNet) or Transformers or incorporate attention mechanisms that help in capturing more complex patterns in the data.

References

[1] Zhao G, Zhang Y, Chu J. A multimodal teacher speech emotion recognition method in the smart classroom. Internet of Things. 2024 Apr 1;25:101069.
[2] Pulido ML, Hernández JB, Ballester MÁ, González CM, Mekyska J, Smékal Z. Alzheimer's disease and automatic speech analysis: a review. Expert systems with applications. 2020 Jul 15;150:113213.
[3] Issa D, Demirci MF, Yazici A. Speech emotion recognition with deep convolutional neural networks. Biomedical Signal Processing and Control. 2020 May 1;59:101894.
[4] Hema C, Marquez FP. Emotional speech recognition using cnn and deep learning techniques. Applied Acoustics. 2023 Aug 1;211:109492.
[5] Zhang C, Xue L. Autoencoder with emotion embedding for speech emotion recognition. IEEE access. 2021 Mar 30;9:51231-41.
[6] Bhangale KB, Kothandaraman M. Speech emotion recognition using the novel PEmoNet (Parallel Emotion Network). Applied Acoustics. 2023 Sep 1;212:109613.
[7] Ahmed MR, Islam S, Islam AM, Shatabda S. An ensemble 1D-CNN-LSTM-GRU model with data augmentation for speech emotion recognition. Expert Systems with Applications. 2023 May 15;218:119633.
[8] Praseetha VM, Joby PP. Speech emotion recognition using data augmentation. International Journal of Speech Technology. 2022 Dec;25(4):783-92.
[9] Mohan M, Dhanalakshmi P, Kumar RS. Speech emotion classification using ensemble models with MFCC. Procedia Computer Science. 2023 Jan 1;218:1857-68.
[10] Eriş FG, Akbal E. Enhancing speech emotion recognition through deep learning and handcrafted feature fusion. Applied Acoustics. 2024 Jun 5;222:110070.
[11] Sefara TJ. The effects of normalisation methods on speech emotion recognition. In2019 International multidisciplinary information technology and engineering conference (IMITEC) 2019 Nov 21 (pp. 1-8). IEEE.
[12] Mushtaq Z, Su SF, Tran QV. Spectral images based environmental sound classification using CNN with meaningful data augmentation. Applied Acoustics. 2021 Jan 15;172:107581.
[13] Zhao J, Mao X, Chen L. Speech emotion recognition using deep 1D & 2D CNN LSTM networks. Biomedical signal processing and control. 2019 Jan 1;47:312-23.
[14] Koduru A, Valiveti HB, Budati AK. Feature extraction algorithms to improve the speech emotion recognition rate. International Journal of Speech Technology. 2020 Mar;23(1):45-55.
[15] Aouani H, Ayed YB. Speech emotion recognition with deep learning. Procedia Computer Science. 2020 Jan 1;176:251-60.
[16] Ancilin J, Milton A. Improved speech emotion recognition with Mel frequency magnitude coefficient. Applied Acoustics. 2021 Aug 1;179:108046.
[17] Chen Q, Huang G. A novel dual attention-based BLSTM with hybrid features in speech emotion recognition. Engineering Applications of Artificial Intelligence. 2021 Jun 1;102:104277.
[18] Andayani F, Theng LB, Tsun MT, Chua C. Hybrid LSTM-transformer model for emotion recognition from speech audio files. IEEE Access. 2022 Mar 31;10:36018-27.
[19] Pham NT, Dang DN, Nguyen ND, Nguyen TT, Nguyen H, Manavalan B, Lim CP, Nguyen SD. Hybrid data augmentation and deep attention-based dilated convolutional-recurrent neural networks for speech emotion recognition. Expert Systems with Applications. 2023 Nov 15;230:120608.
[20] Zhang J, Xing L, Tan Z, Wang H, Wang K. Multi-head attention fusion networks for multi-modal speech emotion recognition. Computers & Industrial Engineering. 2022 Jun 1;168:108078
[21] Singh V, Prasad S. Speech emotion recognition system using gender dependent convolution neural network. Procedia Computer Science. 2023 Jan 1;218:2533-40.
[22] Yao Z, Wang Z, Liu W, Liu Y, Pan J. Speech emotion recognition using fusion of three multi-task learning-based classifiers: HSF-DNN, MS-CNN and LLD-RNN. Speech Communication. 2020 Jun 1;120:11-9.