INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

RESEARCH ARTICLE

Convolutional Neural Network-Based Automatic Speech Emotion Recognition System for Malayalam

V K Muneer[1]*, **K P Mohamed Basheer**[1], **Rizwana Kallooravi Thandil**[1]

1 Department of Computer Science, Sullamussalam Science College, Affiliated to University of Calicut, Kerala, India

OPEN ACCESS

Received: 17-08-2023

Accepted: 23-10-2023

Published: 15-12-2023

Citation: Muneer VK, Basheer KPM, Thandil RK (2023) Convolutional Neural Network-Based Automatic Speech Emotion Recognition System for Malayalam . Indian Journal of Science and Technology 16(46): 4410-4420. https://doi.org/10.17485/IJST/v16i46.2090

* **Corresponding author**.

vkmuneer@gmail.com

Funding: None

Competing Interests: None

Published By Indian Society for Education and Environment (iSee)

ISSN
Print: 0974-6846
Electronic: 0974-5645

Abstract

Objectives: This research work focuses on developing a SER system using CNN and deep learning techniques for a low-resourced Dravidian Indian Language, Malayalam. The importance of speech as a powerful and natural medium of communication, capable of conveying a wide range of information about an individual's mental, behavioral, and emotional characteristics. With the increasing prevalence of human-machine interactions, the study of speech analysis has played a crucial role in bridging the gap between the physical and digital realms. Particularly, the field of emotion identification has gained popularity, as emotions are frequently expressed through speech cues. However, the scarcity of suitable datasets poses a challenge for researchers conducting experiments. **Methods**: In this paper, we address this challenge by employing Long Convolutional Neural Networks (CNN) to effectively recognize sentiments in voice recordings of Malayalam, a low-resource language. We manually construct datasets from audio clips of Malayalam movies and employ the Mel Frequency-Cepstral-Coefficient (MFCC) approach to extract features from the audio signals. **Findings**: By training, classifying, and testing our model using raw speech data from the dataset, the paper proposes a novel approach for recognizing emotions from voice signals processed in Malayalam with an average accuracy of 71%, indicating its ability to correctly predict emotions from vocal utterances in this under-resourced Language. **Novelty:** The novelty of this work lies in its dedication to addressing the challenges of emotion recognition in a low-resource language, the manual creation of datasets, and the successful adaptation of established techniques to a linguistic context where research is relatively scarce. These contributions collectively advance the field of speech emotion recognition and pave the way for further exploration in underrepresented languages.

Keywords: Speech emotion recognition; Malayalam; Natural Language Processing; MFCC; CNN

1 Introduction

Emotions encompass intense feelings or reactions accompanied by physiological changes. They are subjective experiences influenced by various factors such as thoughts, beliefs, and external circumstances. Emotions range from basic instincts like fear and pleasure to complex states like love, anger, and sadness. They significantly impact human behaviour, decision-making, and interpersonal relationships.

Human emotion is a mental state intertwined with paralinguistic characteristics that stem from internal and external events. It involves physiological reactions, external behaviours, and internal feelings. Positive emotions arise when situations align with expectations, while unfavourable events elicit negative emotions. Speech-based emotion recognition can be speaker-dependent or speaker-independent, where emotions are identified based on the voice signals of male and female speakers. Fundamental frequencies, MFCC, and Linear-Prediction-Cepstrum-Coefficient (LPCC) are some of the fundamental aspects of speech processing utilized in this field. In a recent study, the spectrograms of both genuine and fake emotional voices were analyzed, revealing a similar recognition rate between the two scenarios. This highlights the potential for accurately identifying emotions through speech analysis.

The study addresses a critical research gap in the field of speech emotion recognition by focusing on the Malayalam language, which is characterized by limited research resources. The scarcity of studies and datasets for emotional analysis in Malayalam poses a significant challenge for researchers interested in automatic emotion recognition. The lack of prior works in Malayalam and similar low resourced Dravidian languages increases the scope of this research. Recognizing this gap, we aim to provide a solution to the problem of emotion recognition in Malayalam speech, a language with distinct acoustic characteristics and cultural distinctions.

The problem can be defined as follows: Despite the growing importance of speech emotion recognition in various applications, such as human-computer interaction and sentiment analysis, there is a notable lack of resources and research dedicated to the Malayalam language. Existing emotion recognition models primarily cater to well-resourced languages, leaving a gap in our understanding of how emotions are expressed and recognized in this linguistic context. This study seeks to bridge this gap by developing an effective emotion recognition model tailored specifically for Malayalam, addressing the challenges of data scarcity and language-specific features.

Speech plays a vital role in human communication, allowing individuals to sense and interpret emotions through various senses. While this process occurs naturally in humans, it presents challenges when it comes to machines. The classification of emotions is particularly difficult for automatic emotion recognition systems like Speech Emotion Recognition (SER), given the existence of approximately 300 typical emotional states. To address this complexity, the concept of 'palette theory' is employed, drawing parallels with colour theory, where emotions can be decomposed into primary emotions, much like how colours are composed of basic colours. The primary emotions identified are anger, disgust, fear, joy, sadness, and surprise. This approach helps simplify the classification process for SER systems by focusing on these foundational emotional states. The automated SER holds significant importance in various domains. One crucial application lies in call centre conversations, where SER systems can provide valuable insights into customer emotions, aiding in customer service and satisfaction. Additionally, speech-emotion recognition systems find applications in psychiatric diagnosis, intelligent toys, lie detection, and in-car dashboard systems, where the system can assess the driver's mental state for safety purposes. Moreover, in the context of distant learning, an SER system can analyze users' emotions to identify signs of boredom and accordingly adapt the presentation style, and difficulty level, or provide emotional incentives or compromises. These applications highlight the potential of SER systems to enhance various domains by leveraging speech-based emotion recognition capabilities. The flow of model development for this SER is given in Figure 1.

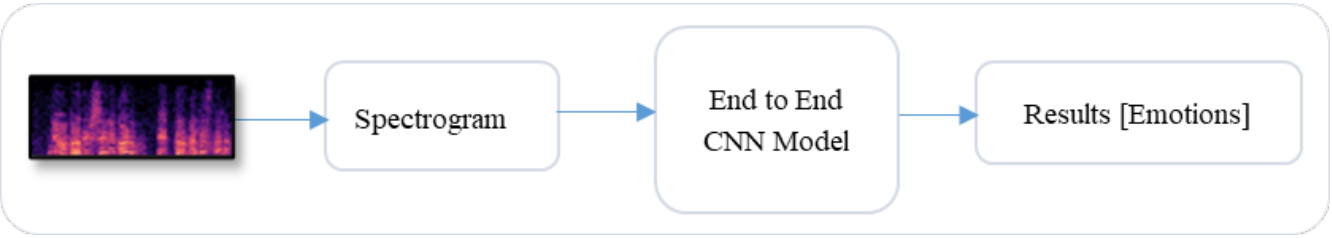

Fig 1. Figure 1: Low-level diagram of SER

Malayalam is a Dravidian language spoken in Kerala, a southern state of India. Voice processing in the Malayalam language presents several challenges. Firstly, Malayalam is considered a low-resource language, meaning there is limited availability of

audio and language resources such as annotated datasets, speech recognition models, and text-to-speech systems. This scarcity hinders the development and training of accurate and robust voice processing algorithms specific to Malayalam. Additionally, the phonetic and acoustic characteristics of Malayalam pose difficulties due to its complex phonological structure, unique phoneme inventory, and regional variations. The lack of standardized pronunciation rules further complicates the accurate recognition and synthesis of Malayalam speech. Furthermore, the limited research and development efforts focused on voice processing in Malayalam add to the challenges, necessitating dedicated efforts to overcome these obstacles and advance voice processing technologies in this language.

The main contributions of this research work are:

1. Creation of a Malayalam sound-emotion-based dataset.
2. Construction of a voice-based embedding to establish associations between emotions and their respective classes.
3. Our proposed CNN architecture utilizes modified kernels and a pooling strategy to extract deep frequency features from speech spectrograms, enabling effective and reliable speech emotion recognition.
4. To improve the performance of baseline SER models, we proposed the utilization of new plain rectangular shape filters for convolutions and pooling, enabling the extraction of deep emotional features from speech spectrograms by leveraging frequencies' pixels with reliable receptive fields.
5. Our proposed SER model achieved an accuracy of 71% when tested on a custom-made dataset. The comparative analysis showcased its superior performance in recognizing emotions in the Malayalam language, making it a simple yet effective system for real-time monitoring of speakers' emotions.

The rapid growth of voice processing aided by machine learning has led to the adoption of various technologies in emotion recognition. Researchers are conducting studies in real-life scenarios to enhance the realism and practicality of these technologies. This approach is causing a significant shift in the accuracy of emotion recognition studies, as new techniques and methodologies are being employed to improve the performance of these systems.

Dutt and Gader (2023) proposed the WaDER method, incorporating an autoencoder, 1D CNN, and LSTM networks for SER, achieving a UA of 81.45% and WA of 81.22% in speaker-dependent experiments [(1)]. Zisad and team (2020) proposed a system to recognize emotions from the speech of individuals with neurological disorders. The proposed model utilizes tonal properties such as MFCCs and the RAVDESS audio speech and song database [(2)]. Mustaqeem. (2021) introduces a novel architecture that involves extracting diverse features from the sound files and utilizing them as inputs for a 1D dilated CNN based on a multi-learning trick approach [(3)].

The approach employed by Abdelhamid et al. (2022) with the help of cascaded layers of feature learning blocks. By extracting high-level features from the log Mel-spectrum of speech samples, the approach effectively captures local correlations and contextual information, resulting in improved emotion recognition performance [(4)]. Lee and Kim (2020) conducted a study on speech emotion recognition, focusing on deep learning techniques including multi-layer perceptron (MLP) and convolutional neural network (CNN). They utilized 1D data extracted from speech files and two-dimensional mel-spectrogram images for training the models and achieved a test accuracy of approximately 60%, with the CNN model showcasing the highest accuracy among the approaches examined [(5)]. Abdulmohsin et al. (2021) introduced a novel statistical feature extraction method for SER which utilizes the variance of feature distribution around the mean, employing multiple degrees of SD on both sides of the mean [(6)]. Issa and Team investigated a model for SER with DCNN by using extracts chromogram, mel-scale spectrogram, MFCC, and Tonnetz representation which resulted in 71.61% accuracy for RAVDESS using 8 classes [(7)].

2 Methodology

The SER system operates akin to a pattern recognition system, consisting of five essential modules: Emotional speech input, Feature extraction, Feature selection, Classification, and Recognized emotional output. These steps are illustrated in Figure 2.

The major modules are:

a. Data Collection and Dataset preparation
b. Data preprocessing
c. Feature Extraction
d. Audio Augmentation
e. Model construction Using CNN
f. Architecture and Training CNN model

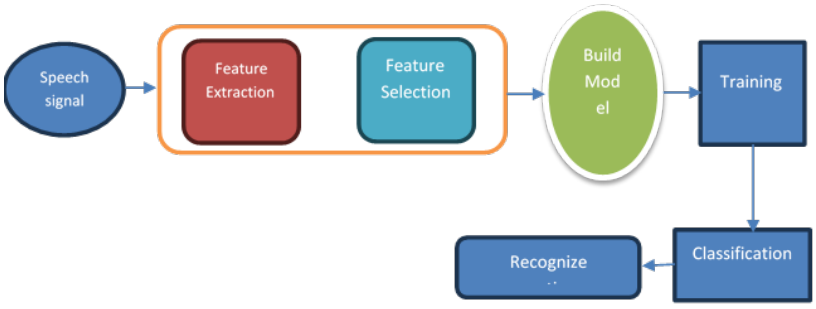

Fig 2. Steps involved in the SER model

2.1 Data collection and dataset preparation

This paper focuses on developing an Automatic SER system for Malayalam speech for Human-Computer Communication. Since there is a lack of available Malayalam language databases for emotion recognition, a new dataset called 'Malayalam_movie_emotions' is constructed for this purpose. The dataset comprises 150 WAV files for each of the seven considered emotions: angry, happy, sad, disgust, surprise, fear, and disgust. These audio files are recorded from popular Malayalam movies using Audacity software, with a sample frequency of 41kHz, and exported as WAV files. To address the issue of background music present in the recorded files, the Spleeter library in Python is utilized for music removal. The dataset is designed to be speaker-independent, including dialogues from both male and female artists, as well as child artists. A total of 25 Malayalam movies were selected, resulting in 1050 distinct sentences in the dataset.

To ensure the availability of suitable data for emotional speech analysis, the authors carefully developed a custom speech corpus under natural recording conditions. This corpus consisted of approximately 1.15 hours of emotional voices, collected from 25 movies in the Malayalam language. The dataset aimed to maintain diversity and representation by including voices from both males and females, covering seven different emotions. Each emotion category in the dataset comprises 150 WAV files. The authors utilized the Spleeter library in Python to address the issue of background music present in the recorded files, effectively removing it. The dataset is designed to be speaker-independent, encompassing dialogues from male and female artists, as well as child artists. Table 1 demonstrates the transcribed text of the Malayalam audio clip and its English translation in each of the sentiment classes.

Table 1. Transcribed text of Malayalam audio clip and it English translation

Sentiments	Malayalam text of audio (sample)	English Translation
Angry	ആരാ നിങ്ങൾക്ക് അധികാരം തന്നത്	Who has given the right to you
Disgust	ഓ.. അഗസ്തി കൂടമല്ല. അസ്ഥികൂടം	Oh. Its not agasthikoodam. It is skelton
Fear	അയ്യോ.. കേട്ടിട്ട് തന്നെ പേടിയാവുന്നു	Oh, I'm scared after hearing it
Happy	സന്തോഷം കൊണ്ട് എനിക്ക് ഇരിക്കാൻ വയ്യേ	I can't sit with happiness
Neutral	അങ്ങനെയും ആവാം	It can be so
sad	ഈശ്വരാ.. ഇങ്ങനെ പരീക്ഷിക്കല്ലേ	God.. Don't try this way
Surprise	ആഹാ.. അത് കേമായല്ലോ	Aha.. that's okay

Data Preprocessing: In this step, the raw audio data is prepared for further processing. This involves tasks such as normalizing the audio, removing background noise, and converting the audio into a suitable format for analysis. During the Feature Extraction phase, relevant features are extracted from the preprocessed audio data. Techniques like Mel Frequency Cepstral Coefficients (MFCCs) are commonly used to capture essential characteristics of the audio signal, such as spectral information and temporal patterns. Audio Augmentation is used to enhance the model's robustness and performance, and audio augmentation techniques are applied. This involves introducing variations like noise addition, pitch shifting, and time stretching to the original audio samples, effectively expanding the dataset. Model Construction Using CN, here, a Convolutional Neural Network (CNN) architecture is designed. CNN is configured to process the extracted features from the audio data, allowing it to learn and recognize patterns that represent different emotional states. The constructed CNN model is trained using the preprocessed and augmented dataset. During training, the model learns to associate the extracted features with their corresponding emotional labels. The training process involves adjusting the model's parameters based on optimization

techniques and loss functions.

2.2 Data preprocessing

After creating the dataset, a data frame is established to store the emotions and their corresponding neural pathways. The characteristics from this data frame will be utilized to train our model. The audio segments in the dataset are separated based on their emotional content, and numerical encoding is applied to represent the emotions. These datasets are then saved to a specified file location. To provide insights into the audio files, wave plots and spectrograms are generated. Wave plots depict the loudness of the audio at different time points, while spectrograms visually represent the frequency spectrum of the sound signal over time. The audio files are all processed with a sampling rate of 44.1 KHz using the 'sr = 44100KHz' parameter in the Librosa library's load function. As for the Deep Convolutional Neural Network (CNN) method, there is no need for preprocessing the data. The raw audio data is directly fed into the neural network model for training. Figure 3 displays the spectrograms generated from the wave files of each one of the seven emotions considered in this work.

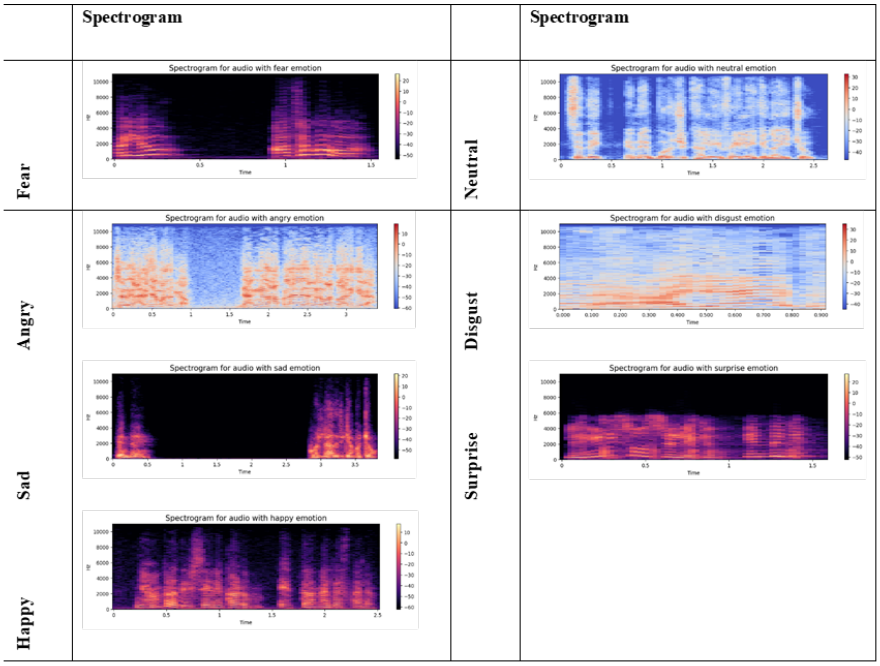

Fig 3. Sample spectrograms of all seven emotions

2.3 Feature extraction

In our study on accented speech recognition, relevant acoustic features are extracted from preprocessed speech data to capture important characteristics such as pitch, tempo, spectral frequencies, and rhythm. We employ three key feature extraction techniques: Mel-frequency cepstral coefficients (MFCCs), Short-term Fourier Transform (STFT), and Tempogram analysis. MFCCs can be computed using the formula MFCCs(t) = DCT(log(E(t) * H(t))), where E(t) represents the magnitude spectrum of the preprocessed speech frame at time t, H(t) is the mel-filterbank matrix, and DCT denotes the Discrete Cosine Transform.

For our study, we focus on the first 13 MFCC coefficients, along with their first and second derivatives. These derivatives provide additional insights into the spectral dynamics of the speech samples, capturing changes in the spectral content over time. The formulas for calculating the first and second derivatives of the MFCC coefficients are as follows:

MFCC'(t) = (MFCC(t+1) - MFCC(t-1))/2

In this process, several features are extracted, including the zero-crossing rate, Chroma STFT, Root Mean Square Value, MelSpectogram, and MFCCs (Mel Frequency Cepstral Coefficients). The feature extraction is performed using the Librosa library, which provides the necessary functionality for extracting these features from the voice data.

2.4 Audio augmentation

Audio augmentation is a technique used to artificially expand the diversity and size of a dataset by applying various transformations to the audio signals. This augmentation process can be applied to the extracted features of emotion-based voice signals obtained from the feature extraction stage. We used augmentations such as Time Stretching: Altering the passage of time involves adjusting the duration of the audio signal without changing its pitch. Noise: Adding noise to the audio signal introduces random variations, simulating different environmental conditions or capturing natural variability in the speech. Pitch Shifting: Changing the pitch of the audio signal involves modifying the frequency content, either raising or lowering the pitch while preserving the original timing. This technique can simulate different vocal characteristics or emotional nuances in speech. Speed Variation: Adjusting the speed of the audio signal involves modifying the playback rate, effectively changing the tempo without affecting the pitch. This technique can simulate variations in speaking rates, emphasizing, or reducing the emotional intensity of the speech. A sample augmented audio signal is represented in Figure 4.

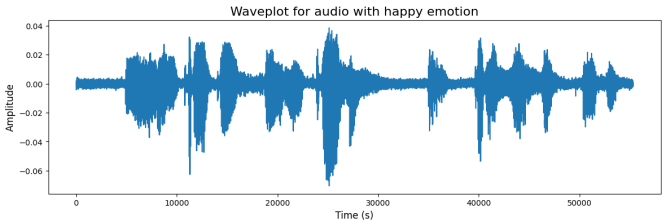

Fig 4. Sample of the augmented audio signal

After augmentation a matrix is formed, that contains 1153x163 features within it. Finally, the augmented data is then fed into the proposed model for emotion prediction.

2.5 Model construction using CNN

The CNN architecture you described consists of four one-dimensional convolutional layers, each followed by a max pooling layer. The number of filters in the convolutional layers progressively decreases from 256 to 128 and then to 64, while the kernel size remains the same at 5 for all layers. The purpose of the convolutional layers is to extract relevant features from the input data. Each filter performs a convolution operation on the input, capturing different patterns or features at different spatial locations. The number of filters determines the complexity and diversity of the learned features. After each convolutional layer, a max pooling layer is applied. Max pooling reduces the spatial dimensions of the feature maps by retaining the maximum value within a pooling window. This helps in capturing the most prominent features while reducing computational complexity and sensitivity to small spatial variations.

The activation function used in the first dense layer is 'ReLU'. ReLU introduces non-linearity to the network by only activating positive values and setting negative values to zero. This allows the network to learn more complex and discriminative representations of the input data. The second dense layer, also known as the output layer, uses the softmax activation function. This is commonly used in multi-class classification tasks, as it allows the model to provide a probability distribution over all possible classes, enabling the selection of the most likely class prediction.

The procedure for developing the SER using CNN is:

Stage 1: Construct the Emotion dataset.

Stage 2: Build the spectrogram dataset that matches the emotion audio.

Stage 3: Prepare the model.

Stage 4: Insert the CNN layers.

Stage 6: Insert the dense layers.

Stage 7: Configure the learning process.

Stage 8: Train the model.

Stage 9: Estimate the model by making predictions.

2.6 Architecture and training CNN model

For training the model, we employed a 1D CNN architecture. The 1D CNN leverages the temporal structure of the audio waveforms, allowing the model to learn patterns and features over time. The key components of the CNN architecture are the

convolutional layer, pooling layer, and fully connected layer. To optimize the model during training, we utilized the ”Adam” optimizer with an initial learning rate of 0.00001. The optimizer helps update the model's parameters to minimize the loss function and the loss function, categorical cross-entropy, as it measures the discrepancy between the predicted and expected outcomes. The training process was conducted over 50 epochs, during which the model iteratively learned from the training data to improve its prediction performance. By leveraging the strengths of the 1D CNN and appropriate optimization techniques, our model aimed to capture the temporal dynamics and relevant features in the audio data for effective speech emotion recognition.

The CNN model architecture provided in the code snippet follows a sequential structure consisting of several layers designed to extract relevant features from input data. Let's break down the architecture step by step:

Input Layer: Input Shape: (x_train.shape$^{(1)}$, 1), where x_train.shape$^{(1)}$ represents the number of input features (MFCC coefficients) and 1 denotes the channel (monophonic audio). Convolutional Layers (4 layers): Conv1D Layer: Each convolutional layer is responsible for learning various features from the input data. It applies filters to capture different patterns in the audio data. Filters: 256 filters in the first layer, 256 filters in the second layer, 128 filters in the third layer, and 64 filters in the fourth layer. Kernel Size: 5 in each layer. Strides: 1 in each layer. Padding: 'same' padding is used to ensure output dimensions match input dimensions. Activation: ReLU (Rectified Linear Unit) activation function is applied element-wise after each convolution operation.

MaxPooling Layers (4 layers): MaxPooling1D Layer: These layers perform downsampling to reduce the spatial dimensions of the data and capture important information. Pool Size: 5 in each layer. Strides: 2 in each layer. Padding: 'same' padding is used to ensure output dimensions match input dimensions. Dropout Layer: Dropout Layer: Regularization technique to prevent overfitting. Rate: 0.2 (20% of neuron outputs are randomly set to zero during training).

Flatten Layer: Flattens the 3D data from the previous layers into a 1D vector for the dense layers. Dense Layers (2 layers): Dense Layer: Fully connected layers that combine features learned by convolutional layers and make final predictions. Units: 32 in the first dense layer, equal to the number of neurons. Activation: ReLU activation function. Dropout Layer: Another dropout layer to further prevent overfitting. Rate: 0.3 (30% of neuron outputs are randomly set to zero during training).

Output Layer: Dense Layer: Final dense layer responsible for outputting class probabilities. Units: Equal to the number of classes (7 in this case, one for each emotion). Activation: Softmax activation function, ensuring the output values represent valid probabilities. The architecture progresses from layers that capture low-level features (convolutional and pooling layers) to higher-level abstractions (dense layers) for emotion recognition. The model is optimized using the Adam optimizer with a categorical cross-entropy loss function. The architecture aims to capture discriminative patterns within the audio data to accurately predict emotions.

3 Results and Discussion

In our model, we incorporated four one-dimensional convolutional layers with varying numbers of filters: 256, 256, 128, and 64. Each convolutional layer utilized a kernel size of 5. Additionally, we included two dense layers in the model architecture. Upon evaluating the model on the test data, we observed an accuracy of approximately 71% as shown in Figure 5.

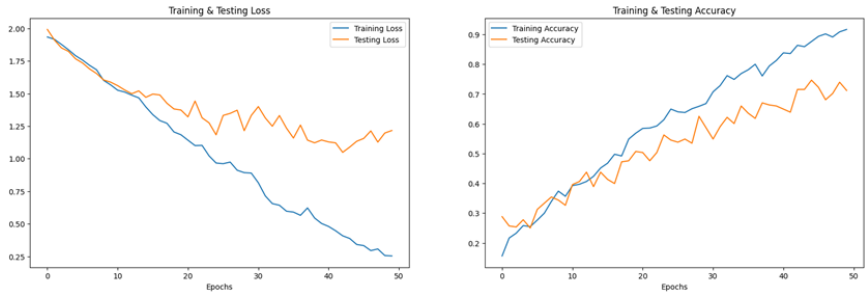

Fig 5. Model Accuracy and Loss: Training and Testing with augmentation

Notably, our model demonstrated higher accuracy in predicting surprise, angry, neutral, and sad emotions. This aligns with expectations, as these emotions tend to exhibit distinct characteristics in terms of pitch, speed, and other audio attributes. Before data augmentation, the model achieved only 57.5% accuracy as given in Figure 6. However, we observed improved performance when augmentation techniques were applied.

The experiment was conducted using a CNN model on the Malayalam dataset aimed to classify and recognize speaker emotions based on their vocal utterances. The evaluation of the experiment considered various metrics such as accuracy,

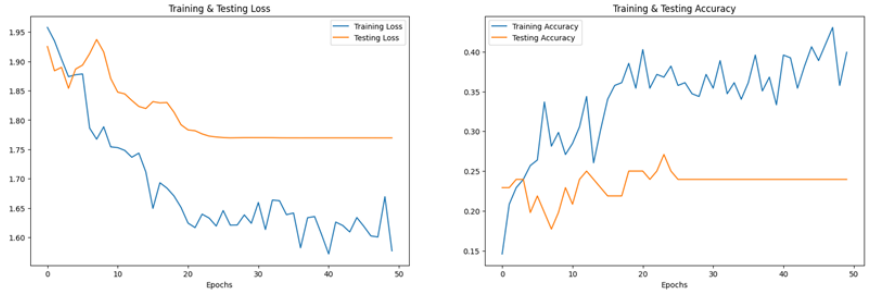

Fig 6. Model Training accuracy and Testing accuracy without Augmentation

precision, recall, and F1-score. The achieved accuracy of 71% indicates that the model correctly predicted the emotions in 71% of the instances. This demonstrates a reasonably high overall correctness of the model's predictions. It is essential to analyze precision, recall, and F1-score along with accuracy to gain deeper insights into the model's ability to identify and capture specific emotions accurately. Figure 7 discusses through confusion matrix.

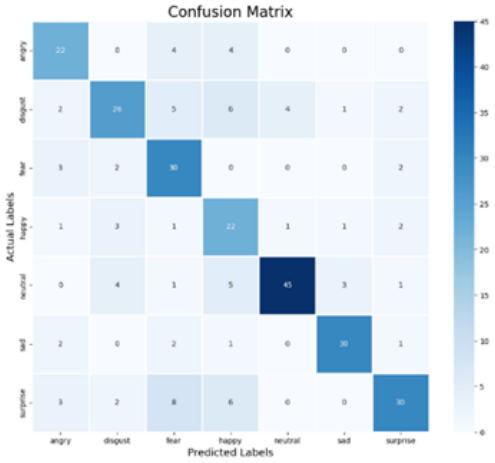

Fig 7. Confusion matrix gives a comparative analysis between the emotion values

Precision, which measures the model's ability to correctly identify instances for each emotion, varied across emotions. The precision values ranged from 50% for the 'happy' emotion to 90% for the 'neutral' emotion. This suggests that the model performed well in identifying 'neutral' emotions but encountered some challenges in accurately predicting the 'happy' emotion. Similarly, the recall values, which indicate the model's ability to find all the relevant instances of each emotion, varied across emotions as well. The recall values ranged from 57% for 'disgust' to 83% for 'sad'. This implies that the model successfully captured a high percentage of instances for the 'sad' emotion but struggled with the 'disgusted' emotion. Table 2 discusses various performance measures of each emotion obtained from the model.

Table 2. Performance measures of different emotions

Emotion	precision	recall	f1-score	support
angry	0.67	0.73	0.7	30
disgust	0.7	0.57	0.63	46
fear	0.59	0.81	0.68	37
happy	0.5	0.71	0.59	31
neutral	0.9	0.76	0.83	59
sad	0.86	0.83	0.85	36
surprise	0.79	0.61	0.69	49
accuracy			0.71	288

Continued on next page

Table 2 continued

macro avg	0.71	0.72	0.71	288
weighted avg	0.74	0.71	0.72	288

The F1-score, which considers the harmonic mean of precision and recall, provides a balanced measure of the model's performance. The F1-score ranged from 59% for 'happy' to 85% for 'sad'. These scores indicate that the model achieved a good balance between precision and recall for the 'sad' emotion but exhibited some room for improvement in accurately predicting the 'happy' emotion.

Table 3 discusses different works carried out, various approaches adopted by different researchers, the dataset used, experimental results and key findings. There is no much prior works observed in speech emotion recognition in Malayalam Language, as it is considered as one the most agglutinative languages in India, moreover highly inflectional and rich in morphology. The observed result in this work is 71% accuracy which is a decent score as far considered these peculiarities of low resourced language.

Table 3. Comparison of similar works with key findings

Ref. No	Approach	Dataset	Key Findings
(8)	GWO-KNN intelligent feature selection method	Arabic Emirati-accented speech database, RAVDESS, SAVEE	Outperformed classical methods for speech emotion recognition using Grey Wolf Optimizer and K-nearest neighbour classifier.
(9)	Speech Emotion Recognition through Hybrid Features and Convolutional Neural Network	Emo-DB, SAVEE, and RAVDESS	CNN model added with MFCCT features outperformed 97%, 93%, and 92% respectively with the dataset
(10)	An Urdu speech corpus for emotion recognition	Corpus collected from 20 subjects	Accuracy of 66.5% with K-nearest neighbours. Removing disgust emotion could produce 76.5% accuracy
(11)	CNN-based speech-emotion recognition system	-	Achieved convincing accuracy of 83.61% on the test dataset, outperforming other methods.
(12)	Comparative study of SER systems using different classifiers and feature extraction methods	Berlin and Spanish databases	Different classifiers achieved accuracies ranging from 83% to 94%, with feature selection techniques improving performance.
(13)	Adaptive Time-Frequency features based on the Fractional Fourier Transform	Berlin EMO-DB, SAVEE, PDREC	The proposed method effectively identified emotional classes with high accuracy on three datasets.
(14)	Ensemble Convolutional Neural Network (ECNN) model	DEAP dataset	Improved accuracy and stability in emotion recognition using multichannel EEG and peripheral physiological signals.
(15)	A two-stage approach using audio features and auto-encoder	-	Better results compared to other SERs
(16)	Dual-level model for SER using MFCC features and mel-spectrograms	IEMOCAP dataset	Significantly outperformed baseline models and achieved comparable results with multimodal models.
(17)	BLSTM and attention model for SER	-	Outperformed in terms of accuracy using speech segment with minimum duration and threshold for silence removal.
(18)	CNN LSTM networks for learning emotion-related features from speech and Clonmel spectrogram	Berlin EmoDB, IEMOCAP	Achieved excellent performance in recognizing speech emotion, outperforming traditional approaches.

Continued on next page

Table 3 continued

[19]	Literature survey on deep learning techniques for SER	EMO-DB, MOCAP	Explored CNN and LSTM architectures, and discussed motivation and experimental findings for speech emotion recognition.
[20]	Feature extraction algorithms for SER	-	Improved recognition rate and accuracy using MFCC, DWT, pitch, energy, ZCR algorithms, and machine learning classifiers.
[21]	Review of deep learning techniques for SER	-	Discusses various deep learning techniques and their architectures for classifying natural emotions, and highlights limitations and challenges.
[22]	Literature review of SER systems	-	Discusses distinct areas of SER, surveys current literature, and highlights challenges in speech emotion recognition.
[23]	RBFN similarity measurement and CNN-BiLSTM model for SER	IEMOCAP, EMO-DB, RAVDESS	Improved recognition accuracy and reduced computational complexity compared to state-of-the-art methods.
[22]	Ensembled model of RNN and DBN for SER in Kannada Language	Manually curated	75% accuracy.
[24]	Ensembled model of SER by parallelizing CNNs with Transformer Encoder.	RAVDESS, IEMOCAP	82.31% accuracy with RAVDEES and 79.42% accuracy for IEMOCAP dataset

Preprocessing Complexity: Preprocessing Malayalam speech discussed in this paper, involves unique complexities due to linguistic nuances and text normalization specific to this language. This complexity sets this work apart from [20], [21], [25], [23], which generally deal with languages that have established preprocessing pipelines.

Accuracy vs. Resources: Despite the linguistic challenges and limited data availability for Malayalam, the proposed model achieves a commendable accuracy of 71%. In comparison, [20] and [21] report higher accuracies, but it's essential to consider the substantial resource advantage they have in terms of larger datasets and feature-rich techniques.

Emphasis on Low-Resource Languages: This work strongly emphasizes addressing emotion recognition in low-resource languages, such as Malayalam, where the lack of annotated data is a significant hurdle. Considering [22], a similar work in Kannada language, which shows SER model produces an accuracy of 75%. In contrast, [20] and [21] mainly focus on languages with more resources, like German, English, and Arabic, which typically have larger datasets available.

Dataset Size: While [24] utilizes RAVDESS and IEMOCAP datasets, [23] uses IEMOCAP, EMO-DB, RAVDESS, established and benchmark datasets, unavailability of dataset compelled to manually construct dataset from Malayalam movies. The process of dataset creation is resource-intensive and adds value to your research in the context of limited data availability for Malayalam.

4 Conclusion

The research work on emotion recognition in Malayalam voice signals using a CNN model has demonstrated promising results. The model achieved an average accuracy of 71%, indicating its ability to correctly predict emotions from vocal utterances. The precision, recall, and F1-score metrics further revealed variations across different emotions, highlighting areas for improvement. While the model performed well for emotions like 'sad' and 'neutral', it encountered challenges in accurately predicting emotions like 'happy' and 'disgust'. Future work should focus on refining the model, exploring data augmentation techniques, and considering alternative feature extraction methods to enhance its performance. Language specialties of Malayalam language, absence of prior work in SER in this low resourced language, Curation on dataset from Movie clips, implementation of SER in Advanced Deep learning techniques for under resourced language and a reasonable experimental accuracy are highlighting points of this research work. Overall, this study contributes to the field of speech emotion recognition in Malayalam and provides insights for further advancements in this area.

Emotion recognition, especially in languages like Malayalam, can be influenced by various factors, including the choice of hyperparameters in CNN model. Hyperparameters such as learning rates, batch sizes, the number of convolutional layers, and the size of kernels can significantly impact the model's performance. Future work could involve curation of larger datasets, fine-tuning for specific emotions, cross cultural considerations and an emphasize on extensive hyperparameter tuning process that might yield better recognition accuracy for specific emotion categories.

References

1) Dutt A, Gader P. Wavelet Multiresolution Analysis Based Speech Emotion Recognition System Using 1D CNN LSTM Networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*. 2023;31:2043–2054. Available from: https://ieeexplore.ieee.org/document/10128692.
2) Zisad SN, Hossain MS, Andersson K. Speech Emotion Recognition in Neurological Disorders Using Convolutional Neural Network. In: International Conference on Brain Informatics, BI 2020;vol. 12241 of Lecture Notes in Computer Science. Springer, Cham. 2020;p. 287–296. Available from: https://doi.org/10.1007/978-3-030-59277-6_26.
3) Mustaqeem, Kwon S. MLT-DNet: Speech emotion recognition using 1D dilated CNN based on multi-learning trick approach. *Expert Systems with Applications*. 2021;167:114177. Available from: https://doi.org/10.1016/j.eswa.2020.114177.
4) Abdelhamid AA, El-Kenawy ESM, Alotaibi B, Amer GM, Abdelkader MY, Ibrahim A, et al. Robust Speech Emotion Recognition Using CNN+LSTM Based on Stochastic Fractal Search Optimization Algorithm. *IEEE Access*. 2022;10:49265–49284. Available from: https://ieeexplore.ieee.org/document/9770097.
5) Lee KH, Kim DH. Design of a Convolutional Neural Network for Speech Emotion Recognition. In: 2020 International Conference on Information and Communication Technology Convergence (ICTC), 21-23 October 2020, Jeju, Korea (South). IEEE. 2020. Available from: https://ieeexplore.ieee.org/document/9289227.
6) Abdulmohsin HA, Wahab HBA, Hossen AMJA. A new proposed statistical feature extraction method in speech emotion recognition. *Computers & Electrical Engineering*. 2021;93:107172. Available from: https://doi.org/10.1016/j.compeleceng.2021.107172.
7) Issa D, Demirci MF, Yazici A. Speech emotion recognition with deep convolutional neural networks. *Biomedical Signal Processing and Control*. 2020;59:101894. Available from: https://doi.org/10.1016/j.bspc.2020.101894.
8) Shahin I, Alomari OA, Nassif AB, Afyouni I, Hashem IA, Elnagar A. An efficient feature selection method for arabic and english speech emotion recognition using Grey Wolf Optimizer. *Applied Acoustics*. 2023;205:109279. Available from: https://doi.org/10.1016/j.apacoust.2023.109279.
9) Alluhaidan AS, Saidani O, Jahangir R, Nauman MA, Neffati OS. Speech Emotion Recognition through Hybrid Features and Convolutional Neural Network. *Applied Sciences*. 2023;13(8):1–15. Available from: https://doi.org/10.3390/app13084750.
10) Asghar A, Sohaib S, Iftikhar S, Shafi M, Fatima K. An Urdu speech corpus for emotion recognition. *PeerJ Computer Science*. 2022;8:1–22. Available from: https://doi.org/10.7717/peerj-cs.954.
11) Qayyum ABA, Arefeen A, Shahnaz C. Convolutional Neural Network (CNN) Based Speech-Emotion Recognition. In: 2019 IEEE International Conference on Signal Processing, Information, Communication & Systems (SPICSCON), 28-30 November 2019, Dhaka, Bangladesh. IEEE. 2020. Available from: https://ieeexplore.ieee.org/document/9065172.
12) Kerkeni L, Serrestou Y, Mbarki M, Raoof K, Mahjoub MA, Cleder C. Automatic Speech Emotion Recognition Using Machine Learning. In: Cano A, editor. Social Media and Machine Learning. IntechOpen. 2019. Available from: https://www.intechopen.com/chapters/65993.
13) Langari S, Marvi H, Zahedi M. Efficient speech emotion recognition using modified feature extraction. *Informatics in Medicine Unlocked*. 2020;20:1–11. Available from: https://doi.org/10.1016/j.imu.2020.100424.
14) Huang H, Hu Z, Wang W, Wu M. Multimodal Emotion Recognition Based on Ensemble Convolutional Neural Network. *IEEE Access*. 2019;8:3265–3271. Available from: https://ieeexplore.ieee.org/document/8941090.
15) Aouani H, Ayed YB. Speech Emotion Recognition with deep learning. *Procedia Computer Science*. 2020;176:251–260. Available from: https://doi.org/10.1016/j.procs.2020.08.027.
16) Wang J, Xue M, Culhane R, Diao E, Ding J, Tarokh V. Speech Emotion Recognition with Dual-Sequence LSTM Architecture. In: ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 04-08 May 2020, Barcelona, Spain. IEEE. 2020. Available from: https://ieeexplore.ieee.org/document/9054629.
17) Atmaja BT, Akagi M. Speech Emotion Recognition Based on Speech Segment Using LSTM with Attention Model. In: 2019 IEEE International Conference on Signals and Systems (ICSigSys), 16-18 July 2019, Bandung, Indonesia. IEEE. 2019. Available from: https://ieeexplore.ieee.org/document/8811080.
18) Zhao J, Mao X, Chen L. Speech emotion recognition using deep 1D & 2D CNN LSTM networks. *Biomedical Signal Processing and Control*. 2019;47:312–323. Available from: https://doi.org/10.1016/j.bspc.2018.08.035.
19) Pandey SK, Shekhawat HS, Prasanna SRM. Deep Learning Techniques for Speech Emotion Recognition: A Review. In: 2019 29th International Conference Radioelektronika (RADIOELEKTRONIKA), 16-18 April 2019, Pardubice, Czech Republic. IEEE. 2019. Available from: https://ieeexplore.ieee.org/document/8733432.
20) Koduru A, Valiveti HB, Budati AK. Feature extraction algorithms to improve the speech emotion recognition rate. *International Journal of Speech Technology*. 2020;23(1):45–55. Available from: https://doi.org/10.1007/s10772-020-09672-4.
21) Khalil RA, Jones EG, Babar MI, Jan T, Zafar MH, Alhussain T. Speech Emotion Recognition Using Deep Learning Techniques: A Review. *IEEE Access*. 2019;7:117327–117345. Available from: https://ieeexplore.ieee.org/document/8805181.
22) Baliga S, Sapna HM, Gowda VY, Patil CM, Arlene A. Kannada Speech Emotion Recognition Using Ensembling Techniques. *IRE Journals* . 2023;6(11):250–255. Available from: https://www.irejournals.com/formatedpaper/1704436.pdf.
23) Mustaqeem, Sajjad M, Kwon S. Clustering-Based Speech Emotion Recognition by Incorporating Learned Features and Deep BiLSTM. *IEEE Access*. 2020;8:79861–79875. Available from: https://ieeexplore.ieee.org/document/9078789.
24) Ullah R, Asif M, Shah WA, Anjam F, Ullah I, Khurshaid T, et al. Speech Emotion Recognition Using Convolution Neural Networks and Multi-Head Convolutional Transformer. *Sensors*. 2023;23(13):1–20. Available from: https://doi.org/10.3390/s23136212.
25) Akçay MB, Oğuz K. Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. *Speech Communication*. 2020;116:56–76. Available from: https://doi.org/10.1016/j.specom.2019.12.001.