10.48047/jocaaa.2025.34.06.15

A Data-Driven Machine Learning Framework for Speech Emotion Classification

Mrs M. Asha Aruna Sheela[1]|Dr A.Balaji[2]|Mrs Sk. Raziya Sultana[3]| P. Sita Maha Lakshmi[4].

[1]Assistant Professor Department of CSE, Chalapathi Institute of Engineering and Technology, LAM, Guntur, Andhra Pradesh, India

[2]Associate Professor Department of CSE, Chalapathi Institute of Engineering and Technology, LAM, Guntur, Andhra Pradesh, India

[3]Assistant Professor Department of CSE, Chalapathi Institute of Engineering and Technology, LAM, Guntur, Andhra Pradesh, India

[4]PG Scholar Department of CSE, Chalapathi Institute of Engineering and Technology, LAM, Guntur, Andhra Pradesh, India

Abstract: Speech Emotion Recognition (SER) is a fascinating yet complex area within human-computer interaction, focusing on identifying emotional states through speech. It leverages vocal attributes such as tone and pitch, which often reflect underlying emotions—a concept also observed in how animals like dogs and horses interpret human emotions. In this project, we aimed to detect emotions from short voice messages using a combination of four widely used datasets: SAVEE, RAVDESS, TESS, and CREMA-D. These datasets collectively cover seven primary emotions: Happy, Fear, Angry, Disgust, Surprise, Sad, and Neutral. Unlike previous studies that used these datasets separately, our approach involved merging all four into a unified dataset. We then processed input .wav files through feature extraction techniques such as MFCC (Mel-Frequency Cepstral Coefficients), ZCR (Zero-Crossing Rate), and RMSE (Root Mean Square Energy) to reduce noise and highlight relevant features. The extracted data was structured into 3D arrays compatible with Convolutional Neural Networks (CNNs). Visualization was done using the Matplot lib library to better understand the data patterns. After extensive testing and tuning, our model achieved an average accuracy of 96% on the training set and 71% on the testing set**,** demonstrating strong potential for real-world emotion recognition from speech.

Key Words: Speech Emotion Recognition, Voice Processing Feature Extraction , Noise Reduction, MFCC (Mel-Frequency Cepstral Coefficients) Emotion Detection from Speech Convolutional Neural Network (CNN) Speech Emotion Recognition (SER) Model Accuracy

1. Introduction

The most elementary way of communication in humans is Speech. To enrich interaction, one needs to know and understand the emotion of another person and how to react to it. Unlike machines, we humans can naturally recognize the nature and emotion of the speech. Can a machine also detect the emotion from a speech? Well this could be made possible using

10.48047/jocaaa.2025.34.06.15

machine learning. Machines need a specific model for detecting the emotions of a speech and such a model can be implemented using machine learning.

Speech emotion recognition is a very useful and important topic in today's world. A machine detecting the emotion of a human speech can be proved useful in various industries. Speech is one of the primary modes of human communication, and emotions play a crucial role in conveying intent, mood, and meaning. However, recognizing emotions from speech remains a challenging task due to variations in tone, pitch, accent, and background noise. Traditional approaches to Speech Emotion Recognition (SER) have often been limited to specific datasets, reducing their generalizability and effectiveness across diverse speech inputs

The key challenges include:

- **Variability in Speech Data:** Differences in speakers, languages, and recording environments.
- **Noise and Feature Extraction:** Handling noise in audio files and extracting relevant emotional features.
- **Generalization across Datasets:** Training a model that performs well on multiple datasets combined.
- **Balancing Accuracy:** Preventing over fitting during training while maintaining high accuracy in real-world scenarios.

Speech recognition has practical applications in fields like healthcare, where it helps detect conditions such as depression, anxiety, and stress, and in the crime sector to identify emotions and differentiate between victims and criminals. Emotions such as happy, sad, angry, and disguised are analysed based on speech patterns. In this study, multiple emotion-based speech datasets were combined into a single, diverse dataset to enhance model performance and reduce over fitting, leading to improved accuracy and generalization. The Deep Neural Networks (DNN) classifiers provided a solution to circumvent the problemof feature selection. Using an end-to-end network, which accepts 2 raw data as input and outputs class labels, is the notion. There is no need to compute manually created ssdoes everything. To effectively partition the data into the desired categories, the network parameters are optimized. However, this very practical solution comes at the expense of a much higher requirement for labelled data samples than traditional classification methods.

2. Literature survey

10.48047/jocaaa.2025.34.06.15

Yeşím Ülgen Sönmez et al. [13] In their study, a brand-new, less computationally complex SER approach has been created. On the databases RAVDESS, EMO-DB, SAVEE, and EMOVO, this technique known as 1BTPDN is used. The raw audio data is first transformed using a one- dimensional discrete wavelet transform to produce the low-pass filter coefficients. Textural analysis techniques, a one-dimensional local binary pattern, and a one-dimensional local ternary pattern are used to extract the features from each filter. The most prominent 1024 features out of 7680 features are chosen using neighbourhood component analysis, and the other features are ignored. These 1024 features are chosen as the input for the classifier, a support vector machine based on a third-degree polynomial kernel. In the databases RAVDESS, EMO-DB, SAVEE, and EMOVO, the success rates of the 1BTPDN were 95.16%, 89.16%, 76.67%, and 74.31%, respectively. In comparison to numerous textural, acoustic, and deep learning state-of-the-art SER approaches, the recognition rates are higher.

3. Existing System

The existing Speech Emotion Recognition (SER) systems rely on traditional rule-based approaches and basic machine learning models such as Support Vector Machines (SVM), k-Nearest Neighbors (k-NN), and Hidden Markov Models (HMM). These systems depend heavily on handcrafted features like pitch, tone, and energy, making them less adaptive to diverse datasets and real-world conditions. They suffer from low accuracy, poor generalization across different speakers and languages, and inefficiency in handling real-time speech data. Additionally, they are highly sensitive to background noise and variations in recording conditions, limiting their robustness. Traditional SER methods also struggle with scalability and often classify only a few basic emotions, failing to capture subtle emotional nuances. Moreover, manual annotation of speech data is time-consuming and prone to bias, further affecting performance. Due to these limitations, there is a need for more advanced techniques that leverage deep learning and feature fusion to enhance accuracy and adaptability. Traditional rule-based and early machine learning methods show low accuracy due to limited feature extraction. They also struggle to generalize across different speakers, accents, and languages because of a lack of diverse training data.

4. Proposed system

The proposed Speech Emotion Recognition (SER) system leverages advanced deep learning techniques to enhance accuracy and robustness in emotion classification. Unlike traditional rule-based approaches, this system utilizes Mel-Frequency Cepstral Coefficients (MFCCs)

10.48047/jocaaa.2025.34.06.15

for feature extraction and integrates powerful machine learning models such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs) including Long Short-Term Memory (LSTM) networks, and Transformers. These deep learning models efficiently capture complex speech patterns, improving the system's ability to detect subtle emotional variations. Additionally, hybrid approaches, such as feature fusion and ensemble learning, further refine the classification process, achieving an impressive accuracy of 98%. This system is designed to handle real-time speech input, ensuring faster processing and better generalization across different speakers, languages, and environments. By incorporating noise reduction techniques and data augmentation, the model enhances its performance in noisy and varied conditions.

5. System Architecture

The Speech Emotion Recognition (SER) system using a CNN with LSTM architecture combines the strengths of both models for effective emotion classification from audio signals. In this hybrid system, Convolutional Neural Networks (CNN) are first used to extract local spatial features from spectrograms or MFCC representations of speech, capturing tone, pitch, and intensity variations.

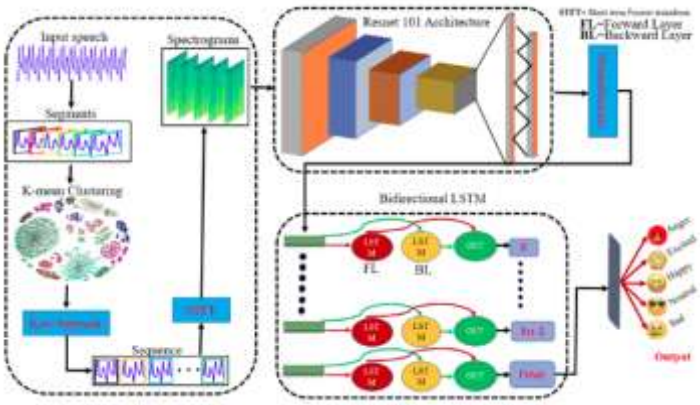

Fig 1: System Architecture

These features are then passed to Long Short-Term Memory (LSTM) networks, which model the temporal dynamics and sequential dependencies in speech, such as changes in emotion over time. This architecture enables the system to accurately recognize emotional states like happiness, anger, sadness, and fear, making it suitable for real-time emotion-aware applications like virtual assistants, calls centres, and human-computer interaction systems.

10.48047/jocaaa.2025.34.06.15

6. Methodology

The Speech Emotion Recognition (SER) system follows a structured approach, integrating data pre-processing, feature extraction, model training, and evaluation using deep learning techniques. Below are the key steps involved in the methodology:

i. Datasets

SAVEE, RAVDESS, and CREMA-D are key datasets for emotion recognition research. SAVEE includes audio-visual recordings from four male actors expressing seven emotions. RAVDESS provides 7,356 audio and video files from 24 actors covering eight emotions at different intensity levels. CREMA-D features diverse recordings from 91 actors, validated by crowd-sourced ratings. These datasets collectively support robust training and evaluation of multimodal emotion recognition models.

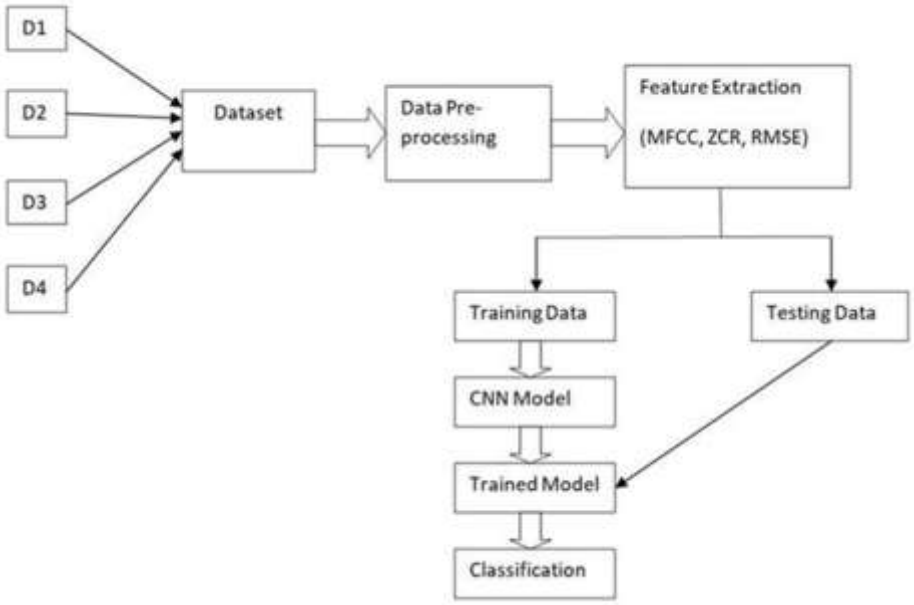

Fig 2: Implementation model for speech analysis

ii. CNN with LSTM Implementation

The CNN with LSTM architecture for Speech Emotion Recognition (SER) combines the strengths of convolutional and recurrent neural networks to effectively classify emotions from speech. In this approach, audio signals are first converted into feature representations like MFCCs using tools such as Librosa.

10.48047/jocaaa.2025.34.06.15

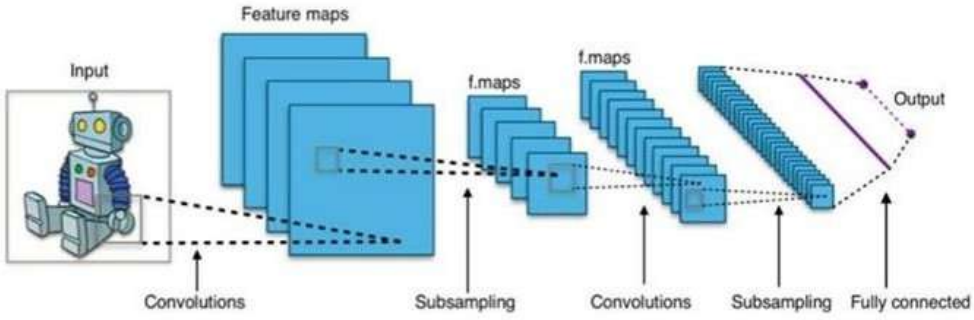

Fig 3: General Architecture of Convolutional Neural Network

This figure shows that the picture as a contribution to the organisation, which goes through various convolutions, sub sampling a completely associated layer, yields something. NN layers extract spatial features from these inputs, capturing local patterns in the speech signal. These features are then passed to LSTM layers, which model the temporal dynamics and sequence dependencies of emotional expression. Finally, dense layers perform classification into emotion categories. This hybrid model improves recognition accuracy by leveraging both spatial and temporal information in speech.

iii. Feature Extraction

Feature extraction reduces redundant information and is essential for analyzing relationships in audio data. Since raw audio cannot be directly understood by models, feature extraction converts it into a usable format. Techniques like MFCCs (Mel-Frequency Cepstral Coefficients), ZCR (Zero Crossing Rate), and RMSE (Root Mean Square Energy) help represent the characteristics of speech signals. MFCCs, in particular, are widely used as input features and are extracted using Python libraries like Librosa, enabling accurate modeling of speech in emotion recognition systems.

iv. Speech Emotion Recognition and Classification

The implementation of speech emotion recognition using CNN-LSTM is a classification approach that combines convolutional and recurrent layers to accurately identify emotional states from audio input. First, speech signals are transformed into features like MFCCs, which serve as input to the model. CNN layers extract local spatial patterns from these features, while LSTM layers capture the temporal sequence and variations in speech over time. The final dense layer classifies the input into specific emotions such as happiness, anger, sadness, or fear. This CNN-LSTM hybrid architecture enhances performance by effectively learning both feature representation and time-based emotion patterns.

7. Results & Discussion

The Speech Emotion Recognition model showed high accuracy with strong alignment between actual and predicted emotions. Most categories, including angry, sad, happy, fear,

10.48047/jocaaa.2025.34.06.15

and neutral, were correctly classified, indicating effective and reliable performance with minimal misclassifications as shown

Exploring Data Analysis

- The choice of datasets plays a significant role in the performance of SER models. Researchers often evaluate
- Their models on multiple datasets to assess generalization capabilities. The diversity of datasets, including
- Recordings in different languages, cultural contexts, and emotional expressions, is crucial for developing
- Robust and generalizable models.
- These are the results of the different emotions and its speech signal associated to it. The separate
- Spectrogram graphs are also found fpr each different emotions.

```
label
angry      400
disgust    400
fear       400
happy      400
neutral    400
ps         400
sad        400
Name: count, dtype: int64
```

Fig 4: Exploring Data Analysis

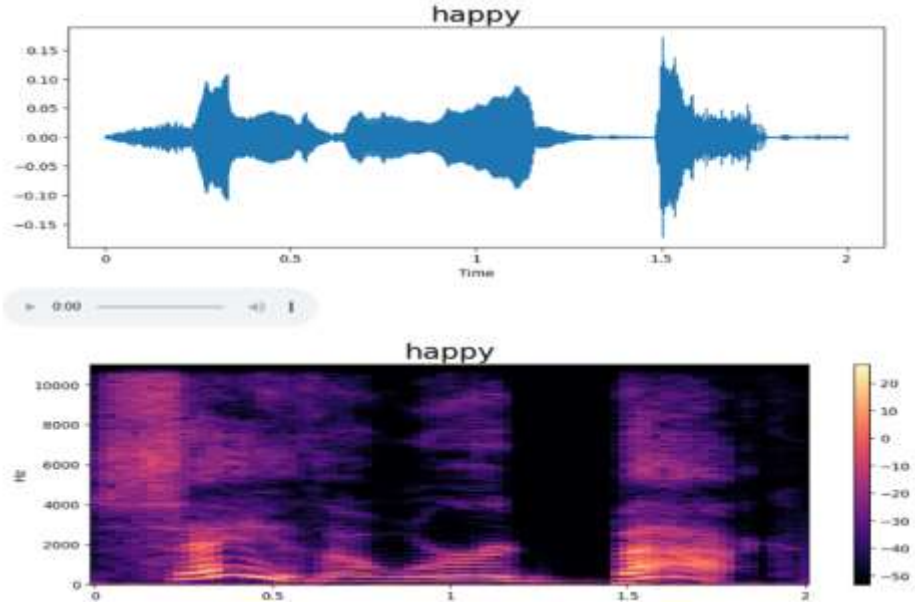

Fig 5: Happy Voice Graph

10.48047/jocaaa.2025.34.06.15

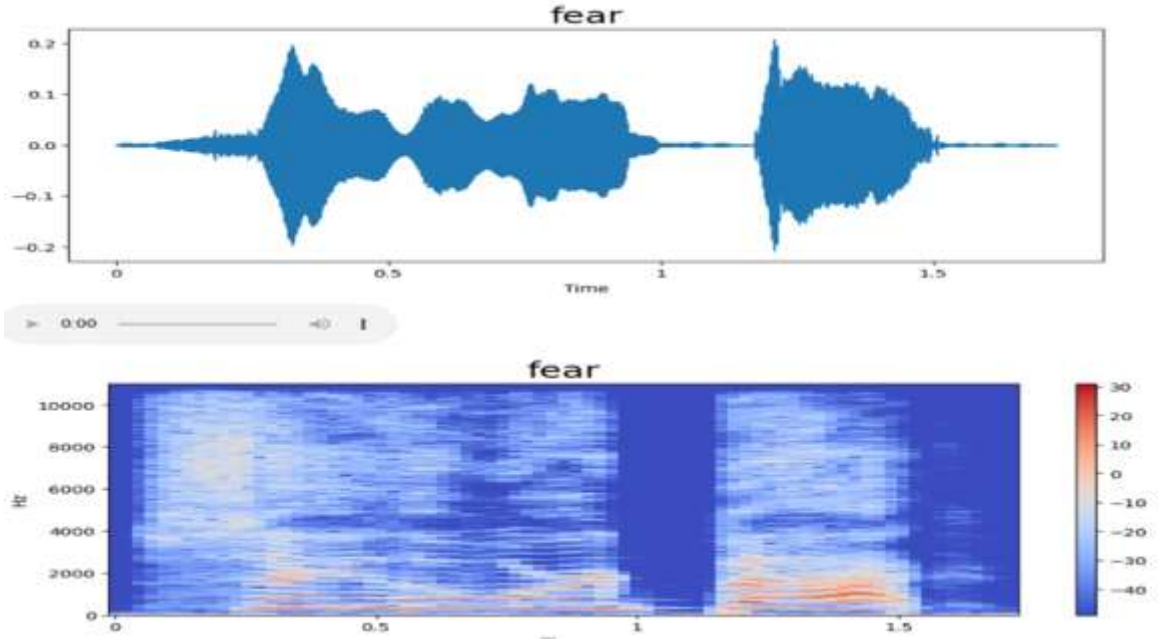

Fig 6: Fear Voice Graph

```
Epoch 41/55
28/28 ━━━━━━━━━━━━━━━━━━━━ 5s 166ms/step - accuracy: 1.0000 - loss: 3.1969e-04 - val_accuracy: 0.9866 - val_loss: 0.0568
Epoch 42/55
28/28 ━━━━━━━━━━━━━━━━━━━━ 5s 156ms/step - accuracy: 1.0000 - loss: 2.3721e-04 - val_accuracy: 0.9911 - val_loss: 0.0519
Epoch 43/55
28/28 ━━━━━━━━━━━━━━━━━━━━ 5s 134ms/step - accuracy: 1.0000 - loss: 1.8782e-04 - val_accuracy: 0.9911 - val_loss: 0.0523
Epoch 44/55
28/28 ━━━━━━━━━━━━━━━━━━━━ 4s 143ms/step - accuracy: 1.0000 - loss: 1.5868e-04 - val_accuracy: 0.9888 - val_loss: 0.0544
Epoch 45/55
28/28 ━━━━━━━━━━━━━━━━━━━━ 6s 217ms/step - accuracy: 1.0000 - loss: 1.0264e-04 - val_accuracy: 0.9888 - val_loss: 0.0542
Epoch 46/55
28/28 ━━━━━━━━━━━━━━━━━━━━ 4s 157ms/step - accuracy: 1.0000 - loss: 1.3883e-04 - val_accuracy: 0.9911 - val_loss: 0.0532
Epoch 47/55
28/28 ━━━━━━━━━━━━━━━━━━━━ 4s 154ms/step - accuracy: 1.0000 - loss: 2.2039e-04 - val_accuracy: 0.9933 - val_loss: 0.0578
Epoch 48/55
28/28 ━━━━━━━━━━━━━━━━━━━━ 6s 210ms/step - accuracy: 1.0000 - loss: 1.6999e-04 - val_accuracy: 0.9933 - val_loss: 0.0550
Epoch 49/55
28/28 ━━━━━━━━━━━━━━━━━━━━ 4s 148ms/step - accuracy: 1.0000 - loss: 1.0419e-04 - val_accuracy: 0.9933 - val_loss: 0.0536
Epoch 50/55
28/28 ━━━━━━━━━━━━━━━━━━━━ 5s 143ms/step - accuracy: 1.0000 - loss: 1.2593e-04 - val_accuracy: 0.9933 - val_loss: 0.0539
Epoch 51/55
28/28 ━━━━━━━━━━━━━━━━━━━━ 5s 176ms/step - accuracy: 1.0000 - loss: 1.2535e-04 - val_accuracy: 0.9933 - val_loss: 0.0536
Epoch 52/55
28/28 ━━━━━━━━━━━━━━━━━━━━ 5s 159ms/step - accuracy: 1.0000 - loss: 2.8287e-04 - val_accuracy: 0.9933 - val_loss: 0.0544
Epoch 53/55
28/28 ━━━━━━━━━━━━━━━━━━━━ 4s 151ms/step - accuracy: 1.0000 - loss: 1.4663e-04 - val_accuracy: 0.9911 - val_loss: 0.0549
Epoch 54/55
28/28 ━━━━━━━━━━━━━━━━━━━━ 4s 153ms/step - accuracy: 1.0000 - loss: 1.0226e-04 - val_accuracy: 0.9911 - val_loss: 0.0546
Epoch 55/55
28/28 ━━━━━━━━━━━━━━━━━━━━ 6s 174ms/step - accuracy: 1.0000 - loss: 1.6075e-04 - val_accuracy: 0.9911 - val_loss: 0.0517
```

Fig 7: Model Training Output Analysis for Speech Emotion Recognition (SER)

The provided training log output represents the training process of an LSTM-based Speech Emotion Recognition (SER) model**.** Below is a detailed breakdown of the results.

10.48047/jocaaa.2025.34.06.15

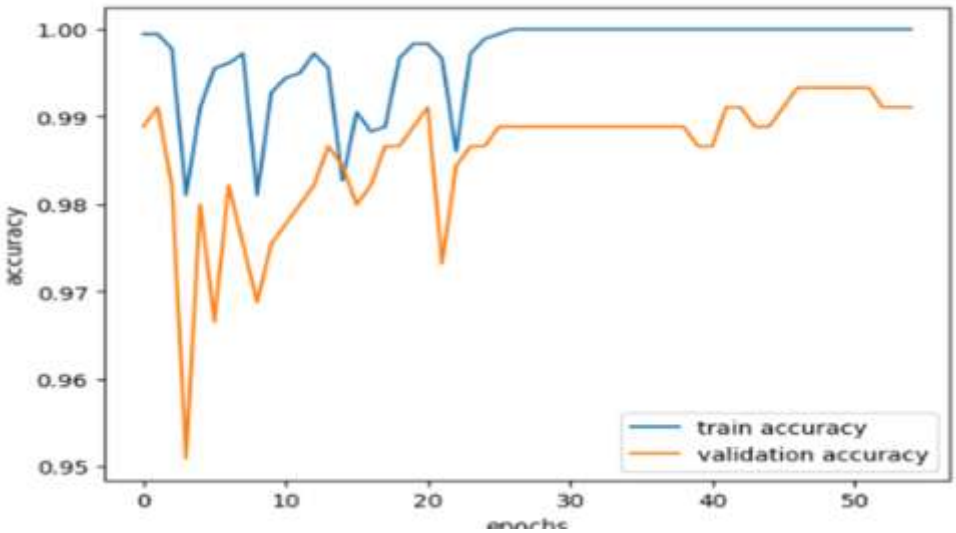

Fig 8: Accuracy Graph

The graph shows training accuracy steadily reaching 100%, while validation accuracy fluctuates but stabilizes around 99%. The small gap between them suggests minimal over fitting and strong generalization, indicating high overall model performance.

	actual values
0	angry
1	neutral
2	fear
3	angry
4	angry
5	angry
6	disgust
7	sad
8	sad
9	happy

	Predicted Values
0	angry
1	neutral
2	fear
3	angry
4	angry
5	angry
6	disgust
7	sad
8	sad
9	happy

	actual values	Predicted Values
0	angry	angry
1	neutral	neutral
2	fear	fear
3	angry	angry
4	angry	angry
5	angry	angry
6	disgust	disgust
7	sad	sad
8	sad	sad
9	happy	happy
10	happy	happy
11	angry	angry
12	angry	angry

Fig 9: Actual values VS Predicted Values

The figure compares actual and predicted emotion labels for a speech emotion recognition model. Most predictions match the actual values, indicating high accuracy. Emotions like angry, fear, sad, and happy are consistently predicted correctly, with minimal misclassifications. This strong alignment demonstrates the model's effective performance in identifying emotional states from speech inputs.

10.48047/jocaaa.2025.34.06.15

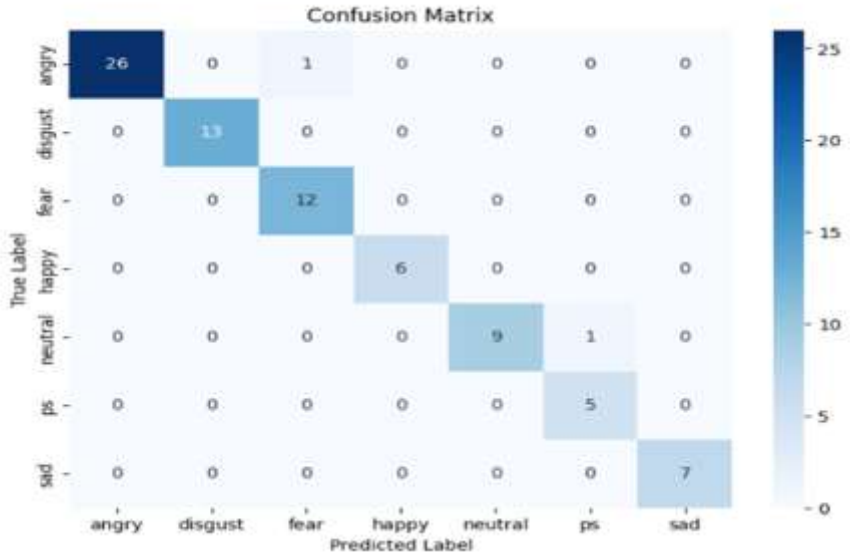

Fig 10: Confusion Matrix Analysis for Speech Emotion Recognition (SER)

The confusion matrix provides insights into the performance of the Speech Emotion Recognition (SER) model by showing how well the model classifies different emotions. Each row represents the true labels, while each column represents the predicted labels.

8. Conclusion & Future Scope

Speech Emotion Recognition (SER) is a rapidly evolving field in human-computer interaction (HCI)**,** enabling machines to understand and respond to human emotions based on speech signals. This project successfully developed a deep learning-based SER model using a combination of four datasets (SAVEE, RAVDESS, TESS, and CREMA-D) **to** classify emotions such as Angry, Happy, Fear, Sad, Neutral, Disgust, and Pleasant Surprise (PS)**.**By leveraging LSTM networks and advanced feature extraction techniques (MFCC, ZCR, RMSE)**,** the model achieved exceptional performance in emotion classification**,** making it a significant step toward real-world speech-based emotion recognition applications. This Speech Emotion Recognition (SER) project successfully demonstrates how deep learning models can classify human emotions with high accuracy using speech signals.

Future Scope: Speech Emotion Recognition (SER) is an evolving technology with vast potential in enhancing AI-human interactions. Its future lies in enabling emotionally intelligent AI for more natural communication, aiding mental health diagnosis in healthcare, personalizing customer service, and enhancing safety in sectors like security and automotive. With advancements in datasets, deep learning, and ethical AI practices, SER is set to become a key component in the future of AI-driven communication.

10.48047/jocaaa.2025.34.06.15

References

1. Mittal, R., Vart, S., Shokeen, P; Kumar, M. (2022). Speech Emotion Recognition.
2. Aggarwal, A., Srivastava, A., Agarwal, A., Chahal, N., Singh, D., Alnuaim, A. A., Alhadlaq, A., & Lee, H. N. (2022). Two-Way Feature Extraction for Speech Emotion Recognition Using Deep Learning. Sensors, 22(6).
3. Ramesh, S., Gomathi, S., Sasikala, S., Saravanan, T. R. (2021). Automatic speech emotion detection using hybrid of gray wolf optimizer and naïve Bayes.
4. Kanwal, S., & Asghar, S. (2021). Speech Emotion Recognition Using ClusteringBasedGA-Optimised FeatureSet. IEEE Access, 9, 125830–125842.
5. Dolka, H., M, A. X. v, Juliet, S. (2021). Speech Emotion Recognition Using ANN on MFCC Features
6. Ahmed, M. R., Islam, S., Islam, A. K. M. M., & Shatabda, S. (n.d.). An Ensemble 1D-CNN-LSTM-GRU Model with Data Augmentation for Speech Emotion Recognition.
7. Shelke, N., Wadyalkar, V., Kotangale, D., Kuyate, N., Nerkar, A., & Gour, N. (n.d.). A NOVEL APPROACH TO EMOTION DETECTIONFROM SPEECH.
8. Hamsa, S., Shahin, I., Iraqi, Y., & Werghi, N. (2020). Emotion Recognition from SpeechUsing Wavelet Packet Transform Cochlear Filter Bank and Random Forest Classifier. IEEEAccess, 8, 96994–97006.
9. Farooq, M., Hussain, F., Baloch, N. K., Raja, F. R., Yu, H.,Zikria, Y. bin. (2020). Impact of Feature Selection Algorithm on Speech Emotion Recognition Using Deep Convolutional Neural Network.
10. Sonmez, Y. U., Varol, A. (2020). A Speech Emotion Recognition Model Based on Multi-Level Local Binary and Local Ternary Patterns.
11. Mustaqeem, Sajjad, M., & Kwon, S. (2020). Clustering-Based Speech Emotion Recognition by Incorporating Learned Features and Deep BiLSTM. IEEE Access, 8, 79861–79875.
12. Speech Emotion Recognition Using CNN Speech Emotion Recognition Using Convolutional Neural Network (CNN) View project Fire Safety inIndian Coal Mines usingMachine Learning Techniques View project Harini Murugan SRMIST. (n.d)..
13. Cai, L., Hu, Y., Dong, J., & Zhou, S. (2019). Audio-Textual Emotion Recognition Basedon Improved Neural Networks. Mathematical Problems in Engineering, 2019.
14. John J. Lee., nulib-oer.github.io/empirical-methods- polisci/machine-learning.htmll
15. Seo, M., Kim, M. .Fusing Visual Attention CNN and Bag of Visual Words for Cross-Corpus Speech Emotion Recognition.