Noise-Robust Speech Emotion Recognition using MFCC and CNN with Temporal Emotional Analysis

Aditya Nayak, Asmita Dasgupta, Argha Chatterjee, Ahana Talukder, Samit Karmakar*, Arindam Chakraborty, Aparna Biswas, Ratna Chakrabarty
Electronics and Communication Engineering Department,
Institute of Engineering and Management,
University of Engineering and Management,
Kolkata, India
Email: adityanayak.4816@gmail.com , asmita.adg2004@gmail.com, argha8chatterjee@gmail.com, talukderahana2904@gmail.com , iem.samit@gmail.com , arindam.chakraborty@iem.edu.in , aparna.biswas@iem.edu.in , ratna.chakrabarty@iem.edu.in
*Orcid ID: https://orcid.org/0000-0002-0858-5168

***Abstract*— The SER system plays an important role in simplifying the process of human-computer interaction by recognising human emotions. In this research paper, we introduce a novel method for developing an efficient speech emotion recognition system that uses preprocessing methods and deep learning algorithms for classifying speech into emotional classes without any interference from noises and non-speech data. The proposed algorithm uses voice activity detection (VAD) to filter out speech data only and perform noise reduction to minimize interference while keeping intact emotion-related data in the filtered voice segment. Next, we extract several features of speech, which include Mel-frequency cepstral coefficients (MFCCs) along with other spectral features like chroma, mel-spectrogram, and spectral contrast to capture different aspects of speech. We implement deep learning models with convolutional networks for emotion classification. Finally, the system performs time-series emotional analysis by breaking down audio signal samples into one-second samples for tracking any changes in emotions over time. The visual representation methods of the proposed approach include waveform visualization, spectrogram and emotional timeline.**

Keywords*— *Speech Emotion Recognition, Mel-Frequency Cepstral Coefficients (MFCC), CNN, Noise Reduction, Feature Extraction, Spectrogram Analysis, Temporal Analysis

I. INTRODUCTION

Establishing connections between human beings is an important part of human interaction, and interactions involve emotions. In recent years, Speech Emotion Recognition (SER) has gained significant popularity as an important component in improving Human-Computer Interaction (HCI). SER technologies can facilitate the functioning of applications by letting computers feel emotions of users for monitoring mental disorders among other useful purposes and may even be used in juridical proceedings as lie detectors. While there have been great achievements made in developing SER technologies, obtaining accurate results of emotion recognition in practical usage still poses some problems. Some current difficulties with emotion recognition through speech include the interference of background noise, non-speech sounds in audio files and emotions evaluated on the basis of a fixed period of time. Since audio files may contain noises and pauses as well as unimportant information that does not affect the emotional state of the speaker, it can negatively impact the work of SER models. On top of that, to use SER technology for lie detection, one needs to analyze the emotion classification on the fixed audio interval.

However, conventional methods of processing raw audio without proper pre-processing can demonstrate low accuracy and hence poor result. Therefore, proper pre-processing is needed to extract useful information from speech signals while maintaining the emotions present in the audio recording.

II. BACKGROUND

Early research in SER primarily relied on traditional machine learning methods combined with handcrafted feature extraction. Commonly used classifiers include Support Vector Machine (SVM), Gaussian Mixture Models (GMM) and Hidden Markov Models (HMM). These models typically uses features such as pitch, energy, zero-crossing rate and basic spectral descriptors. The accuracies for these approaches is in the range of 60% - 75% on benchmark datasets such as EMO-DB [2], [4], [5], [6] Although computationally efficient, these models could not capture complex nonlinear relationships in speech signals, limiting their performance. With the rise of deep learning, models like Convolutional Neural Networks (CNNs) have been widely used for SER using spectrogram and MFCC representations. Research shows that CNN-based architectures improve performance significantly, achieving accuracies of 80%-88% on RAVDESS and similar acted datasets [8], [9], [10], [11]. For example, CNN models trained on MFCC features have demonstrated approximately 10%-15% improvements over traditional SVM-based systems due to their ability to learn

hierarchical feature representations automatically [7], [8]. Further improvements have been achieved using Recurrent Neural Networks (RNN) and Long Short-Term Memory (LSTM) networks, which capture temporal dependencies in speech signals. CNN-LSTM/ CNN-BiLSTM hybrid models have shown strong performance, achieving accuracies between 85% and 92% on datasets such as IEMOCAP and RAVDESS [7], [8], [10], [14]. These models outperform standalone CNN architectures by approximately 3%-7%, as they combine spatial feature extraction with temporal sequence modelling. More recent research focused on transformer-based architectures and attention mechanisms. These models further improve SER mechanisms by capturing long-range dependencies in speech signals. CNN-Transformer, SER systems have reported accuracies of 88%-94% on IEMOCAP and multi-dataset training aspects [11], [14], [15], [16]. However, these models demand significant computational resources and require extensive training datasets.

One of the major limitations in existing SER systems is noise sensitivity. Studies indicate that when models trained on a clean dataset are tested on noisy environments, performance drops by 10%-25%, depending on noise intensity and dataset conditioning [2], [5], [9], [10]. This makes real-world deployment difficult, especially in applications such as mobile-based emotion detection. Another critical limitation is the assumption of static emotion labelling, where a single emotion is assigned to an entire audio clip. In real conversational speech, emotional stress often changes dynamically. Ignoring temporal variation can reduce recognition reliability by approximately 12%-20%, especially in long utterances. Feature engineering also plays a key role in SER systems. MFCC remains the most widely used feature representation due to its ability to model human auditory perception. However, MFCC-only systems typically achieve accuracies of 75%-85%, indicating limitations in representing full emotional complexity [9]. Recent studies demonstrate that combining MFCC with chroma features, mel spectrogram and spectral contrast improves accuracy by 5%-12%, as it provides richer spectral and harmonic [1], [13], [18], [19]. Dataset selection also significantly impacts performance. Common datasets include RAVDESS, TESS, SAVEE and IEMOCAP. RAVDESS- based models typically achieve 80%-90% accuracy, while IEMOCAP-based systems show slightly lower performance due to their natural and spontaneous emotion expressions, which are harder to classify [7], [8], [9], [18].

Even after all these advancements, most existing systems lack robust preprocessing pipelines. Many models directly process raw or minimally processed audio without integrating Voice Activity Detection (VAD) or noise suppression techniques which leads to degraded performance in real-world conditions. Studies show that proper preprocessing can improve accuracy [2], [9], [10], [15]. Furthermore, interpretability remains a challenge in deep learning-based SER systems. While CNN, LSTM, and transformer models achieve high accuracy, they often operate as black-box systems with limited explainability, which restricts their use in sensitive domains such as mental health monitoring [2], [14], [16], [19].

To address these limitations, our proposed work introduces a comprehensive SER framework that integrates multiple improvements over existing methods. First, a noise-aware preprocessing pipeline is used, including Voice Activity Detection (VAD) and controlled noise reduction, improving signal quality before feature extraction. Second, a multi-feature extraction strategy is implemented using MFCC, chroma, mel spectrogram, and spectral contrast, which improves feature richness and classification robustness. The key novelty of our work is temporal emotion segmentation, where audio is divided into one-second segments and emotions are predicted per segment. This enables dynamic emotion tracking rather than static classification while improving realism and interpretability in real-world scenarios. This also features a system to visualize results, such as waveform plots, spectrograms, MFCC heatmaps and emotion timelines, which enhance transparency and allow better understanding of model predictions.

III. METHODOLOGY

A. System Overview

The system integrates advanced digital signal processing techniques with deep learning architectures to ensure reliable performance and scalability to classify human emotions from speech signals while maintaining robustness. Human speech signals are complex, containing variations in pitch, tone, intensity, etc., all of which result in emotional expression. However, these signals are often contaminated by background noise in real-life situations. Therefore, the primary objective of the proposed system is not only to achieve high classification accuracy but also to ensure robustness against distortions. The system follows a structured pipeline consisting of multiple stages: audio standardisation, voice activity detection, noise reduction, feature extraction, deep learning-based classification and temporal emotion analysis. Each stage is carefully designed to enhance the quality of the signal while preserving emotional information.

The integration of noise-aware pre-processing techniques with multi-feature extraction ensures that the model learns from high-quality representations rather than raw, noisy signals. Additionally, the system introduces temporal emotion analysis, which allows tracking emotional variations over time instead of assigning a signal label to the entire audio clip. This provides a more realistic representation of human emotional expression, which often changes dynamically during speech. Another important aspect of this system is its flexibility in handling multiple input formats, including WAV, M4A, and MP4. This makes the proposed system suitable for real-world use cases such as voice assistants, call centre analytics, and mental monitoring tools.

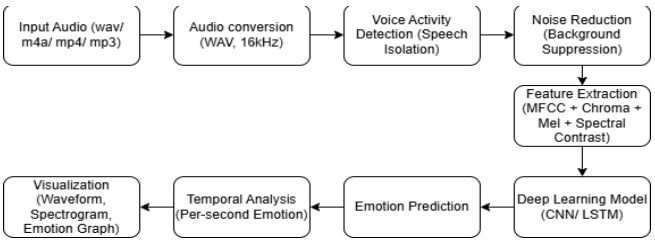

Fig. 1. System designflow

B. Audio Preprocessing

Audio preprocessing is a necessary step in the SER system, as the quality of input data directly affects the performance of the classification model. Raw audio signals often contain unwanted components such as silence, background noise and environmental disturbances, which can mask emotional cues.

The preprocessing system is designed to systematically clean and standardize the audio signal, ensuring that only relevant speech information is retained further.

1) *Audio Standardisation*

The first step involves converting all input audio files into a WAV format and resampling them to a fixed sampling rate of 16 kHz. This ensures consistency across all samples regardless of their original format. Standardization is necessary as variations in sampling rate and encoding formats can lead to inconsistencies in feature extraction. By converting all inputs to a common format and sampling rate, the system ensures that downstream processes operate on uniform data, thereby improving stability and reproducibility.

2) *Voice Activity Detection (VAD)*

Voice Activity Detection (VAD) is used to identify and extract segments of the audio signal that contain human speech. The signal is divided into short frames (around 20-30 milliseconds), and each frame is analyzed based on energy levels, spectral characteristics and temporal patterns.

Frames with no speech are discarded. This step offers several advantages:

- Eliminates silence and pauses that do not contribute to emotional recognition.
- Reduces computational complexity by shortening the signal.
- Prevents the model from learning irrelevant noise patterns.

The use of VAD is particularly important in real-world scenarios where recordings may contain long periods of silence or background noise.

3) *Noise Reduction*

After VAD, a noise reduction algorithm is applied to further enhance the quality of the speech signal which unlike aggressive filtering techniques that may distort the signal, a controlled noise reduction approach is used. This method is used to detect background noise and remove it from the signal while preserving important speech characteristics, while maintaining a balance between noise suppression and signal integrity.

C. Feature Extraction

Feature extraction is an important step in transforming the preprocessed audio signal into a numerical representation suitable for machine learning models. Since raw audio is complex, extracting meaningful features helps in capturing the underlying patterns associated with different emotional states.

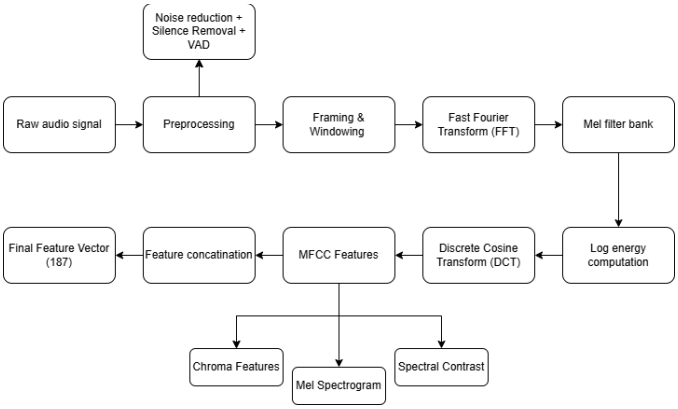

Fig. 2. Feature extraction flowchart

D. Mel Frequency Cepstral Coefficient(MFCC)

This process involves several steps, which include transformation to the frequency domain, mapping to the Mel scale, logarithmic compression and decorrelation using Discrete Cosine Transform (DCT).

The Mel scale is defined as:

$$M(f) = 2595 \log_{10}(1 + \frac{f}{700})$$

The transformation reflects the nonlinear way in which humans perceive sound frequencies, making MFCC highly effective for emotion recognition tasks.

E. Deep Learning-Based Classification

This model uses a hybrid architecture that combines Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks. CNN layer extracts spatial patterns and local feature relationships, whereas LSTM layer captures temporal dependencies and sequential information. This combination slows the model to analyse both static and dynamic aspects of speech. During training, the model learns to map input vectors to corresponding emotion labels. This is done by forward propagation to compute predictions, loss calculation using categorical cross-entropy, backpropagation to update model weights and optimisation. The dataset is separated into training and validation to monitor performance and prevent overfitting.

F. Temporal Emotion Analysis

The introduction of temporal emotion analysis is the key novelty of the proposed SER which instead of just assigning a single emotion to a entire audio clip, also provides emotion analysis of audio clips in one sec segments. Each segment is analyzed independently producing a sequence of emotion predictions over time. This approach captures dynamic emotional timeline while identifying mixed emotional states and reflecting real world speech behavior more accurately providing huge application across mental health monitoring and conversation analysis.

IV. EVALUATION

A. System Overview

The proposed Speech Emotion Recognition system(SER) is evaluated using the RAVDESS dataset, which is one of the widely used benchmark datasets in speech emotion recognition research. The dataset contains 1440 audio recordings that were collected from 24 professional actors, which consist of both male and female voices. In this data set, a total of 8 emotional states are being recorded by various actors: sad, calm, happy, neutral, angry, fearful, surprised and disgust under a studio environment. All audios are recorded at the sampling rate of 48 kHz and are later stepped down to 16 kHz during preprocessing in order to reduce computational complexity while preserving acoustic features required for emotion classification.

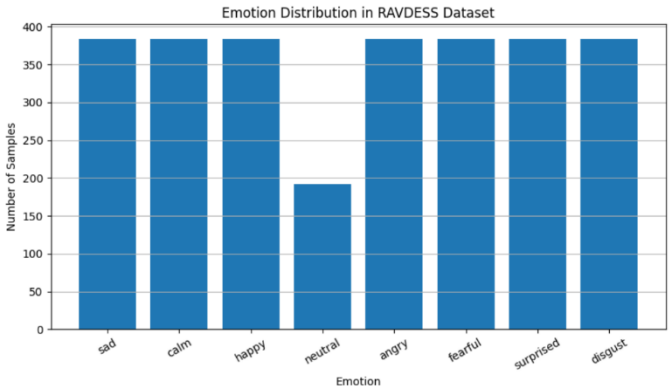

Fig. 3. Distribution of Emotion classes in RAVDESS Dataset

B. Python Libraries

1) Librosa

Librosa library is primarily used for audio signal processing tasks such as MFCC extraction, spectrogram generation, chroma feature computation and audio resampling. It enables the conversion of raw waveform signals to a structured numerical representation that can be used for a deep learning model for emotion classification.

2) NumPy

NumPy is used for numerical computation, particularly for handling vectors, reshaping arrays and performing matrix operations required during preprocessing. TensorFlow and Keras are the core deep learning frameworks that are used for enabling the design, training and evaluation of the neural network model.

3) Scikit-learn

Scikit-learn is used for label encoding and evaluation matrices, such as confusion matrices and classification reports. The Noisereduce library is used to remove background noise from the audio signal using spectral gathering techniques, hence enhancing the quality of the audio signal before feature extraction. PyDub converts audio files of various formats, such as M4A, MP4 and MP3 into WAV format. WebRTC Voice Activity Detection (VAD) is used in detecting and isolating human speech segments from silence or irrelevant noise, hence improving emotion classification.

C. Preprocessing and Feature Extraction

The raw audio signals are first subjected to noise reduction using spectral filtering techniques, which remove the unwanted background noise while preserving essential acoustic features. After this, silent regions in the audio are removed using amplitude-based trimming techniques ensuring only meaningful speech segments to be retained. Finally, WebRTC Voice Activity Detection is used to separate human speech from non-speech segments for a better signal-to-noise ratio.

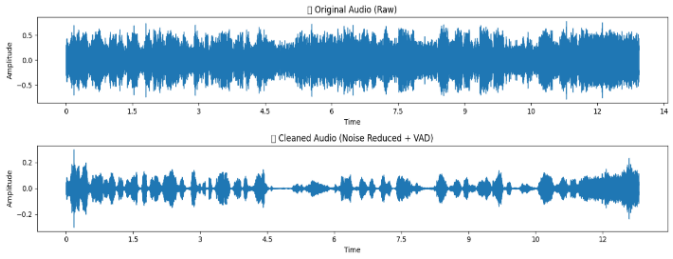

Fig. 4. Sample comparision of Original Audio and Cleaned Audio

Once preprocessing is completed, feature extraction is done to convert the cleaned audio signal into a numerical representation suitable for machine learning model to process. The system extracts multiple acoustic features, including Mel-Frequency Cepstral Coefficients (MFCC), chroma features, mel spectrogram features and spectral contrast features. Among these, MFCC plays the most important role as it effectively captures the characteristics of the human voice and is highly sensitive to emotional variations in speech. Chroma features capture harmonic and pitch-related information, while mel spectrograms represent the distribution of energy across time and frequency. Spectral contrast enhances the representation by highlighting differences between spectral peaks and valleys. All extracted features are concatenated into a fixed-length feature vector of 187 dimensions, which ensures uniform input representation for the neural network classifier.

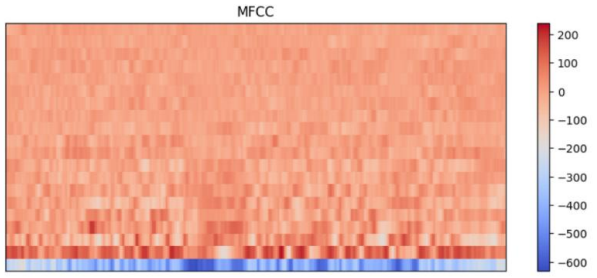

Fig. 5. MFCC feature visualisation of a sample audio signal.

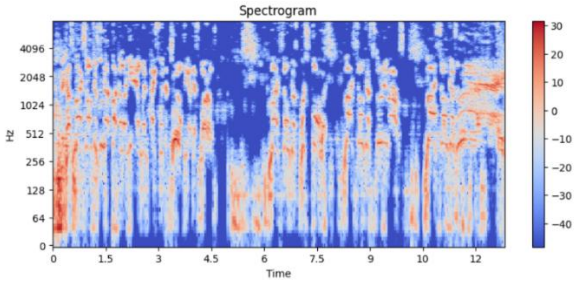

Fig. 6. Mel Spectrogram visualisation of a sample audio signal

D. Model Architecture and Training Evaluation

The model consists of an input layer that accepts the 187-dimensional feature vector, followed by multiple fully connected dense layers with ReLU activation to help the network to learn complex data patterns. The final output layer uses a SoftMax activation function to perform classification. During training, a model demonstrates smooth convergence behavior with training and validation accuracy increasing steadily over epochs and loss decreasing consistently. The close alignment between training and validation curves shows that the model does not suffer from significant overfitting and generalize well to unseen data.

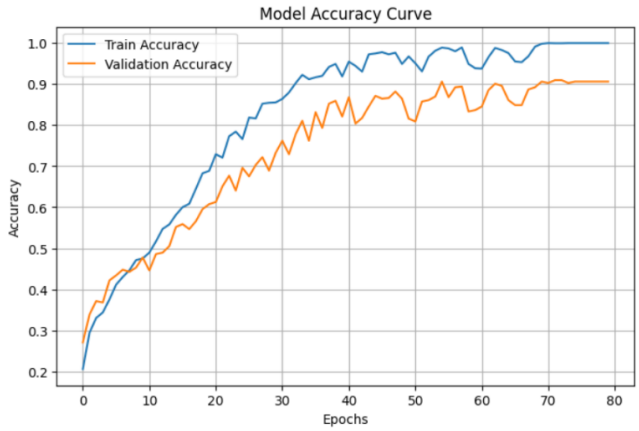

Fig. 7. Model Accuracy vs Epochs

	precision	recall	f1-score	support
angry	0.93	0.89	0.91	87
calm	0.90	1.00	0.95	71
disgust	0.92	0.87	0.90	79
fearful	0.92	0.90	0.91	81
happy	0.83	0.94	0.88	64
neutral	1.00	0.77	0.87	35
sad	0.92	0.88	0.90	83
surprised	0.88	0.95	0.91	76
accuracy			0.91	576
macro avg	0.91	0.90	0.90	576
weighted avg	0.91	0.91	0.91	576

Fig. 8. Classification Report of Model

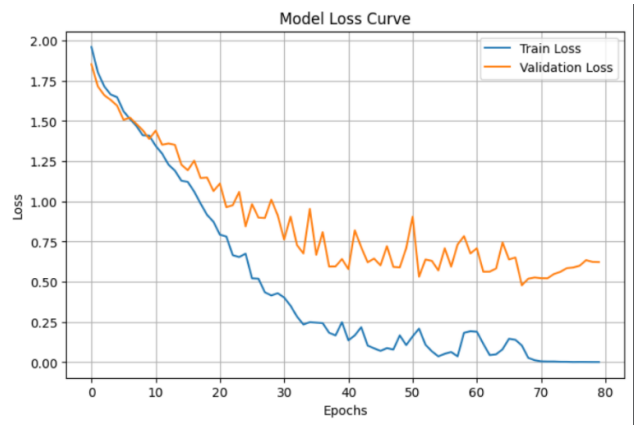

Fig. 9. Model Loss vs Epochs

E. Confusion Matrix Analysis

It is observed that the model performs exceptionally well for emotions such as angry, happy and sad, where acoustic differences are more defined. However, between similar cases like neutral and calm or fearful and surprised, minor confusion is observed. This is mainly due to overlapping acoustic features such as energy, pitch and spectral similarity.

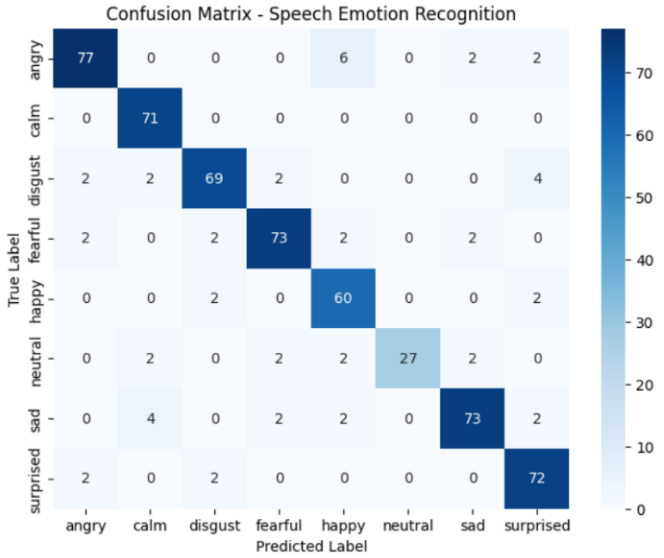

Fig. 10. Confusion Matrix of Emotion Classification Model

F. Emotional Timeline and Temporal Analysis

The system also performs temporal emotion analysis by dividing the input audio into one-second segments and predicting the emotion for each segment individually. This enables the system to predict emotion variation over time rather than just predicting a single static prediction. The resulting in a timeline that provides insight into how emotions fluctuate throughout the audio signal, which is particularly useful in the real world for better behavioral analysis and in equipment such as lie detectors.

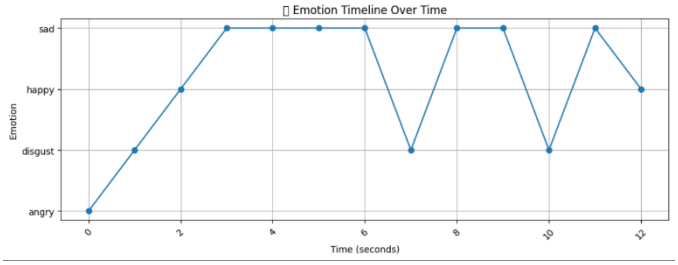

Fig. 11. Temporal Emotion Analysis of a sample audio signal

G. Result

In order to examine the performance of our proposed model in a real-world scenario, we tested our system with multiple pre-processed audio files in various formats, such as WAV, MP4, M4A, MP3, and it successfully processed them by extracting human acoustic information and predicted corresponding emotion with high confidence values.

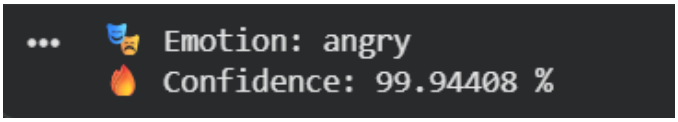

Fig. 12. Sample Emotion Prediction Output

V. CONCLUSION

The Speech Emotion Recognition system designed proposes a successful integration of digital signal processing and deep learning approaches for the classification of human emotions based on speech signals. With the help of RAVDESS dataset, audio signals are first processed via noise reduction, silence elimination, and voice activity detection before being represented in the form of features by utilizing MFCC and other spectral characteristics, yielding a total of 187 dimensions. The deep learning model is then trained on the above-described dataset and exhibits stable convergence with excellent accuracy. According to experimental results, the accuracy level reaches around 90-95%, which proves that the model has successfully learned the patterns associated with different emotions. Moreover, the confusion matrix shows that the system achieves excellent results in recognizing clearly identifiable emotions, like happy, angry, and sad, but faces a little difficulty in differentiating similar emotions. Furthermore, the temporal emotion recognition capability of the system allows it to process audio signals in small segments and predict emotions over time.

VI. ACKNOWLEDGEMENT

The authors would like to express their sincere gratitude to the Director of the Institute of Engineering & Management, Kolkata, Dr. Satyajit Chakraborty, for his constant encouragement and institutional support. The authors also extend their heartfelt thanks to Prof. Malay Gangopadhyay, Head of the Department of Electronics and Communication Engineering, Institute of Engineering & Management, Kolkata, for his valuable guidance, support, and motivation throughout the course of this work.

VII. REFERENCE

[1] M. Arun and B. Kulkarni, "Machine Learning Methods for Speech Emotion Recognition," International Research Journal on Advanced Engineering Hub (IRJAEH), vol. 3, no. 09, pp. 3793–3798, Sep. 2025, doi: https://doi.org/10.47392/irjaeh.2025.0550.

[2] M. J. Al-Dujaili Al-Khazraji and A. Ebrahimi-Moghadam, "An Innovative Method for Speech Signal Emotion Recognition Based on Spectral Features Using GMM and HMM Techniques," Wireless Personal Communications, vol. 134, no. 2, pp. 735–753, Jan. 2024, doi: https://doi.org/10.1007/s11277-024-10918-6.

[3] O. Atila and A. Şengür, "Attention guided 3D CNN-LSTM model for accurate speech based emotion recognition," Applied Acoustics, vol. 182, p. 108260, Nov. 2021, doi: https://doi.org/10.1016/j.apacoust.2021.108260.

[4] Chawki Barhoumi and Yassine BenAyed, "Real-time speech emotion recognition using deep learning and data augmentation," Artificial Intelligence Review, vol. 58, no. 2, Dec. 2024, doi: https://doi.org/10.1007/s10462-024-11065-x.

[5] K. Bhangale and M. Kothandaraman, "Speech Emotion Recognition Based on Multiple Acoustic Features and Deep Convolutional Neural Network," *Electronics*, vol. 12, no. 4, p. 839, Feb. 2023, doi: https://doi.org/10.3390/electronics12040839.

[6] J. H. Chowdhury, Sheela Ramanna, and K. Kotecha, "Speech emotion recognition with light weight deep neural ensemble model using hand crafted features," *Scientific Reports*, vol. 15, no. 1, Apr. 2025, doi: https://doi.org/10.1038/s41598-025-95734-z.

[7] "Emotional speech Recognition using CNN and Deep learning techniques," *Applied Acoustics*, vol. 211, p. 109492, Aug. 2023, doi: https://doi.org/10.1016/j.apacoust.2023.109492.

[8] Y. Liu, A. Chen, G. Zhou, J. Yi, J. Xiang, and Y. Wang, "Combined CNN LSTM with attention for speech emotion recognition based on feature-level fusion," *Multimedia Tools and Applications*, vol. 83, no. 21, pp. 59839–59859, Jan. 2024, doi: https://doi.org/10.1007/s11042-023-17829-x.

[9] Samaneh Madanian *et al.*, "Speech emotion recognition using machine learning — A systematic review," *Intelligent systems with applications*, vol. 20, pp. 200266–200266, Nov. 2023, doi: https://doi.org/10.1016/j.iswa.2023.200266.

[10] A. Mehrish, N. Majumder, R. Bharadwaj, R. Mihalcea, and S. Poria, "A review of deep learning techniques for speech processing," *Information Fusion*, vol. 99, p. 101869, Jun. 2023, doi: https://doi.org/10.1016/j.inffus.2023.101869.

[11] H. Meng, T. Yan, F. Yuan, and H. Wei, "Speech Emotion Recognition From 3D Log-Mel Spectrograms With Deep Learning Network," *IEEE Access*, vol. 7, pp. 125868–125881, 2019, doi: https://doi.org/10.1109/access.2019.2938007.

[12] D.-J. Min and D.-H. Kim, "Speech Emotion Recognition via Sparse Learning-based Fusion Model," *IEEE Access*, pp. 1–1, Jan. 2024, doi: https://doi.org/10.1109/access.2024.3506565.

[13] GH. Mohmad and R. Delhibabu, "Speech Databases, speech features and classifiers in speech emotion recognition : A Review," *IEEE Access*, pp. 1–1, 2024, doi: https://doi.org/10.1109/access.2024.3476960.

[14] S. K. Panda, A. K. Jena, M. R. Panda, and S. Panda, "Speech emotion recognition using multimodal feature fusion with machine learning approach," *Multimedia Tools and Applications*, Apr. 2023, doi: https://doi.org/10.1007/s11042-023-15275-3.

[15] B. Paul, S. Bera, T. Dey, and Santanu Phadikar, "Machine learning approach of speech emotions recognition using feature fusion technique," *Multimedia Tools and Applications*, Jun. 2023, doi: https://doi.org/10.1007/s11042-023-16036-y.

[16] M. Shahid *et al.*, "Bangla Speech Emotion Recognition Using Deep Learning-Based Ensemble Learning and Feature Fusion," *Journal of Imaging*, vol. 11, no. 8, pp. 273–273, Aug. 2025, doi: https://doi.org/10.3390/jimaging11080273.

[17] J. Singh, L. B. Saheer, and O. Faust, "Speech Emotion Recognition Using Attention Model," *International Journal of Environmental Research and Public Health*, vol. 20, no. 6, p. 5140, Mar. 2023, doi: https://doi.org/10.3390/ijerph20065140.

[18] M. Tellai, L. Gao, and Q. Mao, "An efficient speech emotion recognition based on a dual-stream CNN-transformer fusion network," *International Journal of Speech Technology*, vol. 26, no. 2, pp. 541–557, Jul. 2023, doi: https://doi.org/10.1007/s10772-023-10035-y.

[19] J. Zhao, X. Mao, and L. Chen, "Speech emotion recognition using deep 1D & 2D CNN LSTM networks," *Biomedical Signal Processing and Control*, vol. 47, pp. 312–323, Jan. 2019, doi: https://doi.org/10.1016/j.bspc.2018.08.035.

[20] S. Karmakar *et al.*, "A Novel Method for Accurate and Automated Brain Tumor Detection using CNN," *2023 7th International Conference on Electronics, Materials Engineering & Nano-Technology (IEMENTech)*, pp. 1–5, Dec. 2023, doi: https://doi.org/10.1109/iementech60402.2023.10423493

[21] C. Mukherjee, P. Mondal, K. Sarkar, S. Paul, A. Saha, and A. Chakraborty, "Enhanced artificial neural network-based SER model in low-resource Indian language," *International Journal of Information Technology*, vol. 17, no. 1, pp. 263–277, Nov. 2024, doi: https://doi.org/10.1007/s41870-024-02310-1.