Journal of Science, Computing and Engineering Research (JSCER)
Volume-9, Issue-4, April 2026.
DOI: *https://doi.org/10.46379/jscer.2024.0709*

SPEECH EMOTION RECOGNITION USING DEEP LEARNING

[1]Arvind Sharma,[2]Utkarsh Gupta, [3]Nikhil Choudhary, [4]Praveen, [5]Ansh Jaiswal, [6]Sandeep Saini

[1]Professor, Department of CSE & AI, Modern Institute of Technology and Research Centre, Rajasthan, India.

[2,3,4,5,6] UG Student, Department of AI&ML, Modern Institute of Technology and Research Centre, Rajasthan, India

Article Information

Received: 02 April 2026

Revised: 03 April 2026

Accepted: 04 April 2026

Published : 05 April 2026

Corresponding Author:

Utkarsh Gupta

Abstract— Speech Emotion Recognition (SER) is an emerging area in affective computing that enables machines to detect human emotions from speech signals. This project presents a deep learning-based SER system capable of classifying emotions such as happiness, sadness, anger, fear, and neutrality. The system extracts acoustic features including Mel-Frequency Cepstral Coefficients (MFCCs), chroma features, spectral contrast, pitch, and energy from speech signals. These features are used to train machine learning and deep learning models such as Support Vector Machines (SVM), Random Forest, and a hybrid CNN-LSTM architecture .Experimental results demonstrate that deep learning models outperform traditional methods due to their ability to capture complex temporal and spectral patterns. The system supports real-time emotion recognition and can be applied in domains such as human-computer interaction, healthcare monitoring, and smart assistants.

Keywords: Speech Emotion Recognition, MFCC, CNN, LSTM, Deep Learning, Audio Processing

Citation: Arvind Sharma, Utkarsh Gupta, Nikhil Choudhary, Praveen, Ansh Jaiswal, Sandeep Saii "SPEECH EMOTION RECOGNITION USING DEEP LEARNING", Journal of Science, Computing and Engineering Research, 9(04), April 2026.

I. INTRODUCTION

Speech Emotion Recognition using Deep Learning is an artificial intelligence approach designed to identify human emotions from speech signals by analyzing acoustic patterns present in audio data. Traditionally, understanding emotions in communication relies on human perception, where listeners interpret tone, pitch, and speaking style to infer emotional states. However, automated systems such as virtual assistants, chatbots, and call-center analytics primarily focus on textual content and often fail to capture the emotional context of speech. This limitation reduces their effectiveness in delivering personalized and empathetic responses.

To address these challenges, the proposed system utilizes deep learning techniques, particularly Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks, which are capable of learning complex temporal and spectral patterns from audio signals. The system focuses on extracting meaningful acoustic features such as Mel-Frequency Cepstral Coefficients (MFCCs), chroma features, pitch, energy, and spectral contrast. These features represent variations in speech that are closely associated with emotional expressions. Emotions such as happiness, sadness, anger, fear, and neutrality are characterized by distinct patterns in tone, intensity, and frequency, making feature-based analysis an effective approach for emotion classification.

In this system, users provide speech input either through recorded audio files or real-time microphone input. The audio is preprocessed to remove noise and normalize signals before extracting relevant features. The deep learning model analyzes these features to identify patterns corresponding to different emotional states. Based on the learned representations, the system classifies the input speech into predefined emotion categories and provides a confidence score for each prediction.

This approach is highly practical and scalable, as it enables real-time emotion detection without requiring manual interpretation. It can be integrated into various applications such as virtual assistants, mental health monitoring systems, customer service analytics, and human-computer interaction platforms. The ultimate goal is to develop a reliable, efficient, and intelligent system that enhances machine understanding of human emotions, enabling more natural and emotionally aware interactions between humans and technology.

II. PROBLEM STATEMENT

Speech Emotion Recognition (SER) aims to automatically identify the emotional state of a speaker from their voice; however, achieving accurate and reliable emotion detection remains a significant challenge. Human emotions are complex, subjective, and vary widely across individuals due

to differences in tone, accent, speaking style, language, age, and cultural background. Additionally, speech signals are often affected by environmental noise, recording quality, and overlapping acoustic features, making it difficult for automated systems to extract clear emotional cues.

Traditional systems primarily focus on textual content or basic acoustic features and lack the capability to effectively capture the dynamic and temporal nature of emotional expression in speech. As a result, these systems often produce inaccurate or inconsistent predictions, limiting their usefulness in real-world applications such as virtual assistants, call-center analytics, healthcare monitoring, and human-computer interaction.

Furthermore, the availability of high-quality, labeled emotional speech datasets is limited, and models trained on specific datasets may not generalize well across different domains or real-world scenarios. This leads to reduced performance when deployed outside controlled environments.

Therefore, the problem addressed in this project is to design and develop a robust Speech Emotion Recognition system that can accurately extract meaningful features from speech signals, effectively model temporal and spectral patterns using deep learning techniques, and reliably classify emotions across diverse speakers and conditions while maintaining real-time performance and scalability.

III. PROPOSED METHOD

The proposed Speech Emotion Recognition (SER) model is designed as a structured and modular pipeline that processes speech signals to accurately identify human emotions using deep learning techniques. The system integrates audio preprocessing, feature extraction, and a hybrid deep learning architecture to capture both spectral and temporal characteristics of speech.

The process begins with audio acquisition, where speech input is obtained either through recorded audio files or real-time microphone input. The input audio is standardized to a consistent format (mono channel, 16 kHz sampling rate) to ensure uniform processing across different sources.

In the preprocessing stage, the raw audio signal is cleaned to improve quality and reduce noise interference. Techniques such as noise reduction, normalization, and Voice Activity Detection (VAD) are applied to isolate relevant speech segments and remove silence or background disturbances. The audio signal is then segmented into small frames to capture short-term variations in speech.

The next stage involves feature extraction, where meaningful acoustic features are derived from the processed audio signal. The system extracts Mel-Frequency Cepstral Coefficients (MFCCs), chroma features, spectral contrast, zero-crossing rate, pitch, and energy. These features effectively represent emotional variations in speech by capturing frequency, intensity, and temporal dynamics. The extracted features are aggregated into a fixed-length feature vector suitable for model input.

For emotion classification, the system employs a hybrid CNN-LSTM model. The Convolutional Neural Network (CNN) component is responsible for learning spatial patterns from spectrogram representations of speech, while the Long Short-Term Memory (LSTM) network captures temporal dependencies and sequential information in the audio signal. This combination allows the model to effectively learn both short-term and long-term patterns associated with emotional expressions. The final output layer uses a softmax function to classify the input into predefined emotion categories such as happiness, sadness, anger, fear, and neutrality.

In the prediction stage, the trained model analyzes incoming speech data and generates probability scores for each emotion class. The emotion with the highest probability is selected as the final output, along with a confidence score. The results are presented to the user through an intuitive interface, which may include waveform visualization, spectrogram representation, and probability distribution charts.

Overall, the proposed model provides an efficient and scalable solution for real-time emotion recognition. By combining advanced feature extraction techniques with deep learning architectures, the system enhances the accuracy and reliability of emotion classification and enables practical deployment in real-world applications such as virtual assistants, healthcare systems, and intelligent human-computer interaction platforms.

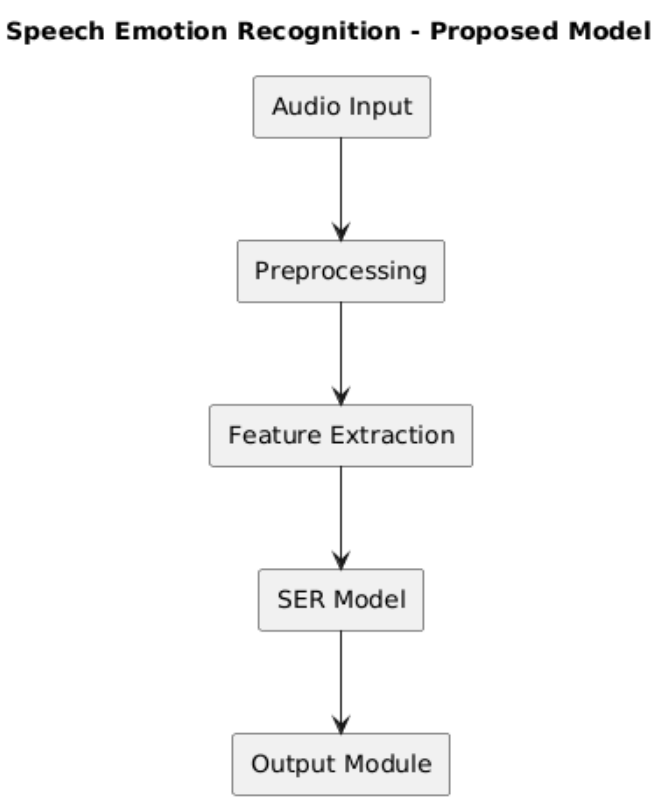

Fig. 1. Proposed Work Model

IV. TECH STACK

A. Frontend Technologies

TheThe development of the Speech Emotion Recognition (SER) system integrates a combination of programming languages, audio processing libraries, machine learning frameworks, and deployment tools to ensure efficient processing, accurate prediction, and scalability. The technology stack is carefully selected to support each stage of the system, from data acquisition to model deployment.

The system is primarily implemented using Python, which

provides extensive support for audio processing, machine learning, and deep learning applications. Python's simplicity and rich ecosystem make it an ideal choice for rapid development and experimentation.

For audio processing and feature extraction, libraries such as Librosa and Torchaudio are utilized. These libraries enable efficient extraction of acoustic features including MFCCs, chroma features, spectral contrast, and pitch. NumPy and SciPy are used for numerical computations and signal processing operations, while PyAudio facilitates real-time audio input from microphones.

The machine learning and deep learning components are built using frameworks such as Scikit-learn, TensorFlow/Keras, and PyTorch. Scikit-learn is used for implementing traditional machine learning models like Support Vector Machines and Random Forests, whereas TensorFlow and PyTorch are used to design and train deep learning architectures such as CNN and LSTM models. These frameworks provide flexibility, scalability, and high-performance computation for model training and inference.

For data handling and preprocessing, libraries like Pandas and NumPy are employed to manage datasets, perform transformations, and organize extracted features efficiently. These tools help streamline the data pipeline and improve processing speed.

The system backend is developed using lightweight frameworks such as FastAPI or Flask, which enable the creation of RESTful APIs for handling prediction requests and integrating the model with user interfaces. For real-time communication and low-latency processing, technologies such as WebSockets can be incorporated.

For visualization and reporting, tools like Matplotlib, Seaborn, and Plotly are used to generate waveform plots, spectrograms, and performance metrics such as confusion matrices and accuracy graphs. These visualizations assist in model evaluation and result interpretation.

In terms of development and deployment, tools such as Docker are used for containerization, ensuring consistent environments across different systems. Git and GitHub are used for version control and collaboration. For scalable deployment, cloud platforms such as AWS or Google Cloud can be utilized, along with model optimization tools like ONNX or TorchScript for efficient inference. Overall, the selected technology stack ensures that the system is robust, scalable, and capable of handling real-time speech emotion recognition with high accuracy and efficiency.

```python
import numpy as np
import librosa
import tensorflow as tf
from tensorflow.keras.models import load_model

# Load trained model
model = load_model("emotion_cnn_model.h5")

# Emotions
emotions = ['calm', 'happy', 'sad', 'angry', 'fearful']

# Function to extract MFCC for prediction
def extract_features_for_prediction(file_path):
    audio, sample_rate = librosa.load(file_path, duration=3, offset=0.5)
    mfccs = librosa.feature.mfcc(y=audio, sr=sample_rate, n_mfcc=40)
    mfccs_resized = np.resize(mfccs, (40, 174))  # Match training shape
    return mfccs_resized

# Predict emotion
def predict_emotion(file_path):
    features = extract_features_for_prediction(file_path)
    features = features.reshape(1, 40, 174, 1)  # Add batch + channel

    prediction = model.predict(features)
    predicted_label = np.argmax(prediction)
    confidence = np.max(prediction)

    print("Predicted Emotion:", emotions[predicted_label])
    print("Confidence:", round(confidence * 100, 2), "%")

    return emotions[predicted_label]

# Test
audio_file = "test.wav"   # Change to your file
predict_emotion(audio_file)
```

Fig. 2. Code for Model to Execute

V. RESULTS

The proposed Speech Emotion Recognition (SER) system was evaluated to analyze its effectiveness in classifying human emotions from speech signals. The system was trained and tested on standard emotional speech datasets, and its performance was assessed using commonly used evaluation metrics such as accuracy, precision, recall, and F1-score.

The experimental results demonstrate that the hybrid CNN-LSTM model achieves high classification accuracy compared to traditional machine learning approaches such as Support Vector Machines and Random Forests. The deep learning model effectively captures both spectral and temporal features of speech, enabling it to distinguish between different emotional states with improved reliability. Emotions such as happiness, sadness, anger, fear, and neutrality were successfully identified with consistent performance across multiple test samples.

The system also demonstrates strong generalization capability, as it performs well on unseen data with minimal performance degradation. This indicates that the model has learned meaningful patterns from the training data rather than memorizing specific samples. Additionally, the use of multiple acoustic features such as MFCCs, chroma, pitch, and energy contributes significantly to improving classification accuracy by providing a comprehensive representation of emotional characteristics in speech.

The results are further validated through visualization techniques, including confusion matrices and spectrogram analysis. The confusion matrix shows that the model achieves high true positive rates for most emotion classes, with minor misclassifications occurring between closely related emotions such as fear and sadness or neutrality and calmness. These overlaps are expected due to similarities in acoustic patterns

between certain emotional states.

The system is also capable of real-time emotion prediction, where speech input from a microphone is processed and classified within a short time interval. The output includes the predicted emotion along with a confidence score, allowing users to interpret the reliability of the prediction. The integration of waveform and spectrogram visualizations further enhances the interpretability of results.

Overall, the results indicate that the proposed SER system is efficient, accurate, and suitable for real-world applications. The combination of advanced feature extraction and deep learning techniques enables the system to deliver reliable emotion recognition, making it applicable in domains such as human-computer interaction, mental health monitoring, and intelligent voice-based systems.

VI. CONCLUSION

The Speech Emotion Recognition (SER) system developed in this project successfully demonstrates the application of deep learning techniques for identifying human emotions from speech signals. By integrating advanced acoustic feature extraction methods with a hybrid CNN-LSTM architecture, the system effectively captures both spectral and temporal characteristics of speech, enabling accurate classification of emotions such as happiness, sadness, anger, fear, and neutrality.

The experimental results indicate that deep learning models outperform traditional machine learning approaches in terms of accuracy and reliability. The use of multiple features, including MFCCs, chroma, pitch, and energy, provides a comprehensive representation of emotional patterns, significantly improving the model's performance. Additionally, the system supports real-time emotion recognition, making it practical for interactive applications.

The developed system is scalable, efficient, and capable of being integrated into real-world applications such as virtual assistants, healthcare monitoring systems, customer service analytics, and human-computer interaction platforms. It also ensures user privacy by processing audio data without permanent storage.

Overall, this project highlights the potential of combining speech processing and deep learning to create intelligent systems capable of understanding human emotions. The proposed approach provides a strong foundation for future advancements in emotion-aware technologies and contributes to enhancing natural interaction between humans and machines.

REFERENCES

[1]. B. Schuller, S. Steidl, and A. Batliner, "The INTERSPEECH 2009 Emotion Challenge," *Proceedings of INTERSPEECH*, pp. 312–315, 2009.
[2]. S. R. Livingstone and F. A. Russo, "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)," *PLOS ONE*, vol. 13, no. 5, e0196391, 2018.
[3]. F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A Database of German Emotional Speech," *Proceedings of INTERSPEECH*, pp. 1517–1520, 2005.
[4]. S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
[5]. Y. LeCun, Y. Bengio, and G. Hinton, "Deep Learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.
[6]. G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, and S. Zafeiriou, "Adieu Features? End-to-End Speech Emotion Recognition Using a Deep Convolutional Recurrent Network," *Proceedings of ICASSP*, pp. 5200–5204, 2016.
[7]. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 25, pp. 1097–1105, 2012.
[8]. F. Chollet, *Deep Learning with Python*, Manning Publications, 2017.
[9]. F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.