2026 Volume: 02 Issue: 04 www.ijrpa.com ISSN 2456-9995 Research Article

International Journal Research Publication Analysis

Page: 01-07

SPEECH EMOTION RECOGNITION USING DEEP LEARNING WITH LSTM NETWORKS

*[1]*Dr. Ramya B. N.*, [2]Smitha U.

[1]Associate Professor

[1,2]Department of Computer Science and Engineering, Jyothy Institute of Technology (JIT), Thataguni, Off Kanakapura Road, Bengaluru – 560082, Karnataka, India.

Article Received: 07 March 2026
Article Revised: 27 March 2026
Published on: 17 April 2026

***Corresponding Author: Dr. Ramya B. N.**
Associate Professor, Department of Computer Science and Engineering, Jyothy Institute of Technology (JIT), Thataguni, Off Kanakapura Road, Bengaluru – 560082, Karnataka, India.

DOI: https://doi-doi.org/101555/ijrpa.8509

ABSTRACT

The growing demand for intelligent human–computer interaction has increased interest in systems that can understand human emotions from speech signals. Speech Emotion Recognition (SER) plays a significant role in applications such as virtual assistants, mental health monitoring, and smart communication systems. This project presents an LSTM-based Speech Emotion Recognition system that analyzes acoustic features extracted from speech audio to classify emotional states. The system utilizes audio feature extraction techniques including Mel Frequency Cepstral Coefficients (MFCC), Chroma features, and Mel Spectrograms to capture important speech characteristics. A Long Short-Term Memory (LSTM) deep learning model is used to learn temporal patterns in speech and classify emotions such as happy, sad, angry, calm, neutral, fearful, disgust, and surprised. The model is trained and evaluated using the RAVDESS emotional speech dataset, achieving an accuracy of approximately 80–85%. This system demonstrates an effective and scalable approach for emotion-aware applications in modern artificial intelligence systems.

KEYWORDS: Speech Emotion Recognition, Deep Learning, LSTM, MFCC, Audio Feature Extraction, RAVDESS Dataset, Emotion Classification.

1. INTRODUCTION

Human communication involves not only spoken words but also emotional expressions that convey important contextual information. Emotions play a crucial role in human interaction,

influencing decision-making, behavior, and understanding between individuals. With the rapid advancement of artificial intelligence and human–computer interaction technologies, there is increasing interest in systems that can automatically detect and interpret human emotions from speech signals. Speech Emotion Recognition (SER) is an emerging field that focuses on identifying emotional states such as happiness, sadness, anger, fear, and neutrality from speech audio. Traditional methods for emotion detection relied on manual analysis and rule-based approaches, which were often limited in accuracy and scalability. However, recent developments in machine learning and deep learning have significantly improved the ability to analyze complex speech patterns and detect emotional cues automatically. Deep learning models, particularly Recurrent Neural Networks (RNN) and Long Short-Term Memory (LSTM) networks, have shown promising results in capturing temporal dependencies in speech signals, making them well-suited for emotion recognition tasks.

In a speech emotion recognition system, audio signals must first be transformed into meaningful numerical features that represent the characteristics of speech. Commonly used acoustic features include Mel Frequency Cepstral Coefficients (MFCC), Chroma features, and Mel Spectrograms, which capture frequency, pitch, and spectral information from speech signals. These features provide essential information that helps machine learning models identify patterns associated with different emotional states.

This project proposes a Speech Emotion Recognition system using a deep learning model based on Long Short-Term Memory (LSTM) networks. The system extracts multiple audio features from speech recordings and uses them to train an LSTM model capable of classifying emotions accurately. The RAVDESS emotional speech dataset is used for training and evaluation of the model. The proposed approach aims to improve the performance of emotion classification and demonstrate the effectiveness of deep learning techniques in analysing emotional information from speech.

2. Structure of the System

The proposed Speech Emotion Recognition system is designed to identify human emotions from speech signals using deep learning techniques. The system consists of several stages including audio input processing, feature extraction, data preprocessing, model training, and emotion classification. Initially, speech audio files from the RAVDESS dataset are provided as input to the system. Important acoustic features such as Mel Frequency Cepstral Coefficients (MFCC), Chroma features, and Mel Spectrogram features are extracted from the audio signals using the Librosa library. These features capture essential characteristics of

speech that help in identifying emotional patterns. The extracted features are then normalized and prepared for training using appropriate preprocessing technique. A Long Short-Term Memory (LSTM) neural network is used to learn temporal patterns present in speech data and classify different emotional states. After training, the model can analyze new speech inputs and predict emotions such as happy, sad, angry, calm, neutral, fearful, disgust, and surprised. This structured approach enables the system to effectively recognize emotions from speech signals and support emotion-aware intelligent applications

3. Literature Overview

Traditional approaches to emotion recognition from speech relied on manual feature extraction and classical machine learning algorithms such as Support Vector Machines (SVM), Decision Trees, and Hidden Markov Models (HMM). While these methods achieved moderate success, they often struggled to capture the complex temporal patterns present in speech signals. With the advancement of deep learning, Speech Emotion Recognition systems have significantly improved by leveraging neural network architectures capable of learning sequential dependencies in audio data. Prior research mainly focuses on extracting acoustic features such as Mel Frequency Cepstral Coefficients (MFCC), Chroma features, and Mel Spectrograms to represent the characteristics of speech signals. These features are then used as inputs for machine learning or deep learning models to classify emotional states. Recent studies have demonstrated that Recurrent Neural Networks (RNN) and Long Short-Term Memory (LSTM) networks are particularly effective for speech emotion recognition because they can model temporal relationships in speech. While many existing models achieve high classification accuracy using benchmark datasets such as RAVDESS and EMO-DB, they often focus primarily on model performance rather than practical implementation. Our proposed system builds upon these approaches by combining multiple acoustic features with an LSTM-based deep learning model to improve emotion classification accuracy and provide a scalable solution for emotion-aware intelligent systems.

4. METHODOLOGY

The proposed Speech Emotion Recognition system is designed to identify emotional states from speech signals using deep learning techniques. The methodology involves several stages including data collection, feature extraction, data preprocessing, model training, and emotion prediction. Initially, speech audio samples are collected from the RAVDESS emotional speech dataset, which contains recordings representing various emotions such as

happy, sad, angry, calm, neutral, fearful, disgust, and surprised. Each audio file is processed using the Librosa library to extract important acoustic features including Mel Frequency Cepstral Coefficients (MFCC), Chroma features, and Mel Spectrogram features. These features capture the spectral, pitch, and frequency characteristics of speech signals and help the model learn emotional patterns effectively.

After feature extraction, the data is re-processed by converting the extracted features into numerical arrays and normalizing them using StandardScaler to improve model performance. The emotion labels are encoded using Label Encoder to convert categorical emotion values into numerical format suitable for training. The dataset is then divided into training and testing sets to evaluate the performance of the model.

The classification model used in this project is based on a Long Short-Term Memory (LSTM) neural network, which is a type of recurrent neural network capable of learning temporal dependencies in sequential data such as speech signals. The architecture consists of multiple LSTM layers followed by dropout layers to reduce overfitting and dense layers for emotion classification. The model is trained using the Adam optimizer and categorical cross-entropy loss function. During training, the network learns patterns from the extracted audio features to accurately classify different emotional states. Once the training process is complete, the model can analyze new speech inputs and predict the corresponding emotion based on learned patterns.

5. Implementation

The implementation of the proposed Speech Emotion Recognition system is carried out using the Python programming language along with several machine learning and audio processing libraries. Libraries such as Librosa are used for extracting important audio features including Mel Frequency Cepstral Coefficients (MFCC), Chroma features, and Mel Spectrogram features from speech recordings. These extracted features are converted into numerical arrays and normalized to improve the performance of the learning model. The emotion labels associated with each audio file are encoded using appropriate preprocessing techniques. The system uses a Long Short-Term Memory (LSTM) based deep learning model implemented using the TensorFlow/Keras framework. The model is trained on the RAVDESS emotional speech dataset and learns patterns in speech signals corresponding to different emotional states. After training, the system can analyze new speech inputs and classify the corresponding emotion such as happy, sad, angry, calm, neutral, fearful, disgust, or surprised

with improved accuracy.

6. RESULTS & DISCUSSION

The performance of the proposed Speech Emotion Recognition system was evaluated using the RAVDESS emotional speech dataset. The dataset was divided into training and testing sets to measure the effectiveness of the model in classifying different emotions. The Long Short-Term Memory (LSTM) based deep learning model was trained using extracted audio features such as Mel Frequency Cepstral Coefficients (MFCC), Chroma features, and Mel Spectrogram features. After the training process, the model was tested on unseen speech samples to evaluate its prediction capability. The experimental results show that the proposed model achieves an accuracy of approximately **80–85%** in identifying emotions such as happy, sad, angry, calm, neutral, fearful, disgust, and surprised. The results demonstrate that combining multiple audio features with an LSTM-based architecture.

7. CONCLUSION

The proposed Speech Emotion Recognition system demonstrates the effectiveness of deep learning techniques in identifying human emotions from speech signals. By extracting important acoustic features such as **Mel Frequency Cepstral Coefficients (MFCC), Chroma features, and Mel Spectrogram features**, the system captures essential speech characteristics that help in distinguishing emotional patterns. These features are used to train a **Long Short- Term Memory (LSTM)** neural network capable of learning temporal dependencies present in speech data. The model was trained and evaluated using the

RAVDESS emotional speech dataset, achieving an accuracy of approximately **80–85%** in classifying emotions such as happy, sad, angry, calm, neutral, fearful, disgust, and surprised. The results indicate that combining multiple audio features with an LSTM-based architecture improves the performance of emotion recognition systems. Such systems have practical applications in areas such as **human–computer interaction, virtual assistants, mental health monitoring, customer service analysis, and intelligent communication systems**. In the future, the system can be further improved by using larger datasets, hybrid deep learning models such as **CNN-LSTM architectures**, and real-time speech emotion detection for interactive applications.

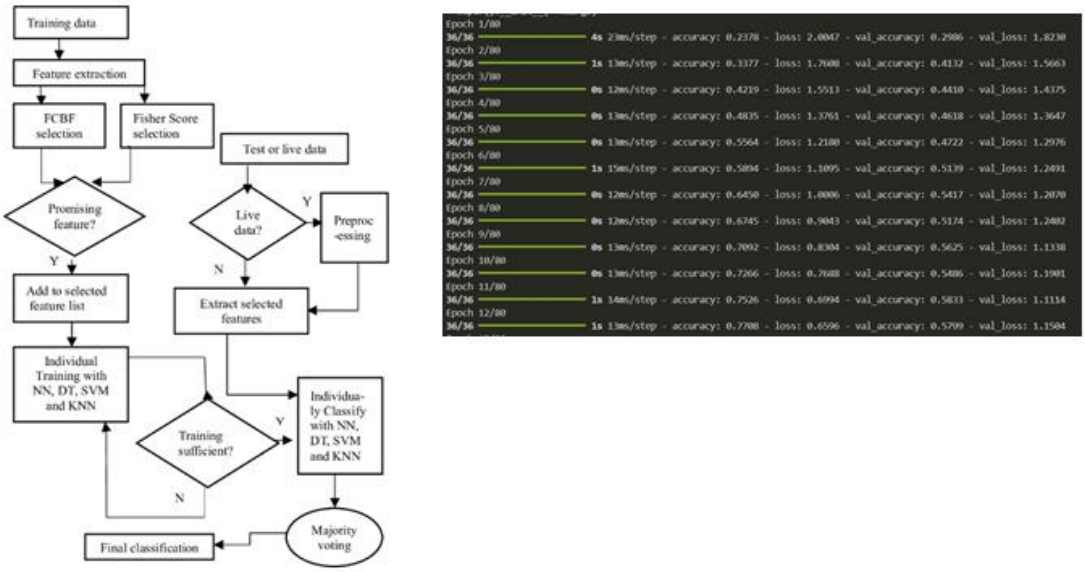

Fig. 1 – Speech Emotion Recognition System:(a) Training performance of the LSTM model showing accuracy and loss across epochs; (b) Overall workflow illustrating speech input, feature extraction, model training, and emotion classification.

8. FUTURE SCOPE

The proposed Speech Emotion Recognition system demonstrates promising results using deep learning techniques and audio feature extraction methods. However, there are several areas where the system can be further improved in the future. One possible enhancement is the use of larger and more diverse speech datasets to improve the generalization capability of the model. Incorporating advanced deep learning architectures such as Convolutional Neural Networks (CNN), Transformer models, or hybrid CNN-LSTM models may further improve emotion classification accuracy. In addition, real-time speech emotion detection can be implemented using microphone input for interactive applications. The system can also be integrated into intelligent virtual assistants, mental health monitoring systems, and customer service platforms to better understand user emotions. Future research may also focus on multilingual emotion recognition and multimodal emotion detection by combining speech with facial expressions or text analysis.

REFERENCES

1. Livingstone, S. R., and F. A. Russo, **"The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS),"** PLOS ONE, vol. 13, no. 5, 2018. [Online]. Available: https://zenodo.org/record/1188976. [Accessed: Dec. 4, 2025].

2. Python Software Foundation, **"Python Documentation,"** 2023. [Online]. Available: https://docs.python.org/. [Accessed: Dec. 4, 2025].
3. F. Pedregosa et al., **"Scikit-learn: Machine Learning in Python,"** J. Mach. Learn. Res., vol. 12, pp. 2825–2830, 2011. [Online]. Available: https://scikit-learn.org/. [Accessed: Dec. 4, 2025].
4. B. McFee et al., **"Librosa: Audio and Music Signal Analysis in Python,"** Proceedings of the Python in Science Conference (SciPy), 2015. [Online]. Available: https://librosa.org/. [Accessed: Dec. 4, 2025].
5. S. Hochreiter and J. Schmid Huber, **"Long Short-Term Memory,"** Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997. [Online]. Available: https://www.bioinf.jku.at/publications/older/2604.pdf. [Accessed: Dec. 4, 2025].