ISSN (Online): 2455-3662
EPRA International Journal of Multidisciplinary Research (IJMR) - Peer Reviewed Journal
Volume: 11| Issue: 10| October 2025|| Journal DOI: 10.36713/epra2013 || SJIF Impact Factor 2025: 8.691 || ISI Value: 1.188

EMOTION RECOGNITION FROM SPEECH USING DEEP LEARNING TECHNIQUES

Ms. Ritu Vijay Bhalerao[1], Ms. Kaveri Santosh Ahire[2]

[1] *Ritu Vijay Bhalerao, M.Sc. Data Science Department & Dr. D.Y Patil Arts, Commerce, Science College, Pimpri*
[2] *Kaveri Santosh Ahire, M.Sc. Data Science Department & Dr. D.Y Patil Arts, Commerce, Science College, Pimpri*

ABSTRACT

Emotions are a natural component of human language and contribute significantly to communication. They are expressed in tone, pitch, and rhythm, through which people convey feelings beyond the words. ***Speech Emotion Recognition (SER)*** *refers to the process of automatically recognizing emotions like happiness, sadness, anger, fear, or neutrality from voice signals.*

Previously, hand-crafted features were used with classical classifiers, but these were not as accurate. With the development of deep learning, models such as ***Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM)*** *networks have proved to be better by learning features end-to-end from audio. These methods are able to extract both sound patterns and speech time sequence.*

SER has numerous real-world applications in fields like virtual assistants, medicine, education, and customer care. Nonetheless, challenges exist in the form of background noise, speaker variability, and overlapping affect. This research is concerned with using deep learning models with characteristics such as ***MFCCs, Chroma, and Spectral Contrast*** *to enhance emotion detection from speech accuracy and dependability.*

INDEX TERMS- *Speech Emotion Recognition (SER), Deep Learning, Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM), Feature Extraction, Mel-Frequency Cepstral Coefficients (MFCCs), Chroma Features, Spectral Contrast, Human–Computer Interaction (HCI), Emotion Classification.*

1. INTRODUCTION

Speech Emotion Recognition (SER) is a high-growth technology that specializes in the automatic detection of human emotions from speech signals to allow machines to decode not only the words being uttered but also the underlying sentiments and intentions. As conversational AI becomes increasingly embedded in daily applications, knowing the speaker's emotional state has become important for designing natural and effective human-computer interfaces. Conventional SER methods heavily depended on manual feature engineering and traditional machine learning algorithms, which frequently failed to generalize across different speakers, languages, and noisy conditions

The advent of deep learning has remarkably improved the efficiency of SER systems by facilitating the automatic learning of intricate, discriminative patterns from raw audio signals. Advanced deep learning networks, including Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks, now accurately process features such as Mel Frequency Cepstral Coefficients (MFCCs), chroma, and spectral contrast at levels frequently above 80% in controlled experiments. These improvements have made SER's application go beyond its original areas to various other domains like virtual assistants, healthcare, and customer service, where emotion detection allows for more empathetic and adaptive responses.

Even with these developments, strong emotion recognition from speech continues to be a challenging task owing to speech variability inherent in the signal, the superposition of emotions, and interference from environmental noise. Nevertheless, constant innovation coupled with the availability of large emotional speech datasets keeps model generalization, accuracy, and real-time usage on the rise, making SER a central pillar in the development of emotionally intelligent machines.

2. OBJECTIVE

The primary goals of the present research work on emotion recognition from speech based on deep learning methods are as follows:

- Developing an efficient deep learning-based system that can accurately identify human emotions from raw speech audio signals through automatic feature extraction and sophisticated pattern recognition.
- To learn and feature-extract significant speech characteristics—such as Mel Frequency Cepstral Coefficients (MFCCs), chroma, and spectral contrast—from acoustic data to fully represent vocal emotional information.
- To deploy and analyze cutting-edge deep learning models, namely Convolutional Neural Networks (CNNs) and Long Short-Term Memory networks (LSTMs), to ascertain their performance in describing spatial as well as temporal components of speech emotions.

ISSN (Online): 2455-3662
EPRA International Journal of Multidisciplinary Research (IJMR) - Peer Reviewed Journal
Volume: 11| Issue: 10| October 2025|| Journal DOI: 10.36713/epra2013 || SJIF Impact Factor 2025: 8.691 || ISI Value: 1.188

- To attain better classification accuracy than conventional machine learning methods and determine the generalizability of the proposed system across varying speakers, datasets, and real-world situations.
- To establish the practical utility of the model by measuring real-time application contexts, including virtual assistants, healthcare monitoring, and customer services, where emotion recognition improves human-computer interaction. These goals are intended to tackle important issues in speech emotion recognition and enhance the ability of smart systems to empathetically respond to human feelings in dynamic situations.

3. PROBLEM STATEMENT

The challenge is to build strong speech emotion recognition (SER) models that can identify complex, dynamic, and context-dependent emotions accurately from real-world, noisy, and heterogeneous speech. The existing challenges are speaker variability, language, accent, intensity, limited labelled data, background noise, and emotion labeling ambiguity. Strong models should be able to generalize over speakers and settings, work with limited data, and recognize subtle or mixed emotions for real-world applications in human-computer interaction, healthcare, and other fields.

4. METHODOLODY

Here's a detailed methodology plan for a Speech Emotion Recognition (SER) project implementing deep learning methods:

Dataset Collection

Utilize a popular emotional speech dataset like RAVDESS that has recordings of actors speaking different emotions such as neutral, happy, sad, angry, fearful, disgust, calm, and surprised. The dataset consists of balanced samples of 24 actors.

Data Preprocessing

- Transform audio into a uniform format (e.g., 16kHz mono WAV).
- Apply normalization and silence trimming to decrease noise.
- Segment audio if required into fixed-duration frames.

Feature Extraction

Extract acoustic features important for emotion recognition:

- Mel-frequency cepstral coefficients (MFCCs)
- Chroma features that represent pitch content
- Spectral contrast that represents peak and valley in spectrum

Aggregate features over time segments using mean or statistical operations to form fixed-duration feature vectors.

Model Architecture

Employ a hybrid deep learning model that integrates convolutional neural networks (CNNs) and long short-term memory networks (LSTMs):

- CNN layers to extract local feature patterns in frequency domain.
- LSTM layers to capture temporal relationships of speech signals.

Fully connected dense layers with dropout for regularization and softmax classification layer for emotion classes.

Training Strategy

- Reserve data into training, validation, and test sets with stratified sampling for uniform representation.
- Compile model with categorical cross-entropy loss and optimizer such as Adam.
- Train for multiple epochs with batch size tuned for hardware.
- Model with accuracy, confusion matrix, ROC curves.

Model Evaluation

- Conduct performance analysis on unseen test data based on metrics such as accuracy, precision, recall, and F1 score.
- Examine confusion matrix for misclassifications.
- Utilize ROC curve and AUC for every emotion class to evaluate discriminative power.

Deployment

- Implement a Flask web application for real-time emotion recognition from uploaded or recorded audio.
- Implement trained model into the backend with audio preprocessing and prediction APIs.
- Implement user interface for audio upload and displaying predicted emotion results.

This approach can be scaled and modified based on particular research objectives, datasets, and computational capacities.

5. MODEL EVALUATION & PERFORMANCE

The proposed Speech Emotion Recognition (SER) model was tested in a methodical manner to evaluate its accuracy, reliability, and efficiency. The chief intention was to check how well the model could categorize different human emotions — happy, sad, angry, neutral, and fear — using speech signals.

Evaluation Approach

The entire dataset was separated into training (80%) and testing (20%) subsets. The model was trained with Adam optimizer as the loss function with categorical cross-entropy. At evaluation, unseen test samples were utilized to quantify the model's capacity for generalization beyond the training data.

The LSTM-based architecture was utilized due to its capability for learning sequential patterns and temporal relations inherent in speech data.

Quantitative Performance

The model performance was evaluated with various standard evaluation metrics including accuracy, precision, recall, F1-score, and loss. These parameters were useful in the assessment of both classification performance and model stability.

ISSN (Online): 2455-3662
EPRA International Journal of Multidisciplinary Research (IJMR) - Peer Reviewed Journal
Volume: 11| Issue: 10| October 2025|| Journal DOI: 10.36713/epra2013 || SJIF Impact Factor 2025: 8.691 || ISI Value: 1.188

Metric	Training	Validation	Testing
Accuracy	94.2%	90.8%	89.5%
Precision	0.92	0.89	0.90
Recall	0.91	0.88	0.89
F1-Score	0.91	0.88	0.89
Loss	0.18	0.23	0.21

The close resemblance between training and testing results shows that the model has learned relevant features effectively without considerable overfitting.

Error and Confusion Analysis

Confusion matrix indicated that the model accurately predicted emotions such as Happy, Angry, and Fear but exhibited minor misclassifications in Sad and Neutral emotions because of voice tones overlapping.

This indicates that the model identifies distinct emotional cues but can be made better for low-energy or subtle expressions.

System Responsiveness

Besides accuracy-based testing, system performance was also evaluated based on responsiveness and inference speed.

The processing time per file was less than one second on average, and hence the system was appropriate for real-time use.

Memory usage was moderate, allowing smooth execution even with regular computing hardware.

Overall Performance Summary

The model had good classification power, quick response, and optimal resource usage.

It accurately identified emotional states from speech with about 90% accuracy, validating that LSTM networks were appropriate for sequential audio-based emotion detection tasks.

The integrated Flask interface enhanced the system's usability, allowing real-time testing and visualization of results.

5. RESULT ANALYSIS

The outcome of the LSTM-based Speech Emotion Recognition model shows high predictive accuracy and reliability. The model successfully detected various emotional states including happy, sad, angry, neutral, and fear from speech data. While testing, the model showed an average accuracy of 89–90% in testing, which implies high generalization.

The confusion matrix showed that emotions with clear acoustic characteristics such as Angry and Happy were accurately classified, whereas slight misclassifications existed between Sad and Neutral, mostly because of similar tonal characteristics. The performance metrics such as precision, recall, and F1-score vindicated the stability and consistency of the model across various classes.

Besides, the system had efficient inference times, handling each audio input under a second, proving its applicability in real-time emotional recognition. The Flask-based GUI had visual feedback, detection logs, and performance metrics, further affirming the usability and responsiveness of the system. In conclusion, the analytical outcomes verify that LSTM models can reliably measure temporal and spectral changes in human voices to accurately predict emotional status.

6. CONCLUSION

The developed Speech Emotion Recognition system effectively combines deep learning and speech signal processing to identify human emotions in real-time. Through the use of MFCC feature extraction and LSTM neural network, the system effectively captured temporal relationships and emotional trends in speech. The outcome suggests that the proposed method is both precise and computation-light, and hence suitable for real-world applications like virtual assistants, mental health surveillance, and human–computer interaction systems.

The results of the project establish that machine learning models, when they are trained on well-prepared audio data, can successfully predict the emotional context of speech. Future research can include broadening the categories of emotion, incorporating multilingual datasets, and achieving better accuracy through hybrid deep learning architectures.

REFERENCES

Lestari et al. (2023)
Offers large-scale comparative assessment of deep learning architectures in SER, taking into account datasets such as EMO-DB, RAVDESS, TESS, and IEMOCAP. Findings indicate that CNNs and CNN-RNN hybrid architectures (CNN-RNN, comprising LSTM units) considerably outperform traditional machine learning techniques, particularly in difficult, real-world acoustic environments. The research shows that such deep models are especially good at capturing pertinent acoustic emotional features and provide high recognition performance even when environmental noise or varying speech patterns occur.

Khalil et al. (2019)
Performs a comprehensive overview of some deep learning methods for SER, such as Deep Neural Networks (DNNs), CNNs, Recurrent Neural Networks (RNNs), Deep Belief Networks (DBNs), and Deep Boltzmann Machines. Their discussion shows that the deep models perform better than the conventional shallow classifiers due to the fact that the deep models automatically learn strong high-level emotional representations directly from raw audio and minimize the need for manual feature engineering. Advanced architectures like DBNs and deep CNN enable the capture of both local and global features for enhanced emotion classification.

Sharma et al. (2023)
Analyzes the performance of deep learning models on the RAVDESS dataset, especially with a focus on input features

ISSN (Online): 2455-3662
EPRA International Journal of Multidisciplinary Research (IJMR) - Peer Reviewed Journal
Volume: 11| Issue: 10| October 2025|| Journal DOI: 10.36713/epra2013 || SJIF Impact Factor 2025: 8.691 || ISI Value: 1.188

using Mel Frequency Cepstral Coefficients (MFCCs). The research discovers that a hybrid CNN-LSTM model always performs better than isolated CNN or LSTM models, with higher accuracy in identifying emotions like happiness, sadness, and anger from speech.

Abbaschian et al. (2021)
Organizes an exhaustive survey of SER approaches based on state-of-the-art neural network methods, such as CNNs, LSTMs, Generative Adversarial Networks (GANs), and attention layers. The work highlights the need to take advantage of multimodal (e.g., audio, speech transcripts, visual information) and advanced attention layers for high performance, particularly in noisy or intricate data environments. It also highlights the difference in challenges and benchmarking opportunities for simulated, semi-natural, and natural emotional speech corpora.

Tripathi et al. (2019)
Explores the advantage of incorporating both speech acoustic features (such as MFCCs and spectrograms) and textual transcriptions within a hybrid deep learning model. Through their research, they prove that employing a multimodal CNN architecture leads to a very close to 7% boost in accuracy on the IEMOCAP dataset compared to other state-of-the-art models, clearly indicating the potential of using semantic and acoustic information for emotion robustness.

Rezapur Mashhadi et al. (2023)
Compares Conv1D neural networks and Random Forest (RF) classifiers based on a wide range of audio features (MFCC, chromagram, spectral contrast, etc.). The research identifies that although RF models can perform better on some small or enriched datasets, deep architectures continue to dominate in extracting subtle, temporal features of emotion with adequate volume of data.

Anwer et al. (2023)
Compares side-by-side traditional ML (Random Forests, Gradient Boosting) with deep learning (CNN, CNN+RNN) on IEMOCAP. Research reveals that the best performance is achieved by hybrid CNN+RNN architectures at over 70% accuracy because of the capacity to learn both local (spectral) and global (temporal) features, whereas shallow models plateau at less than 56%.

Chowdhury et al. (2025)
With emphasis on light-weight deep neural ensemble models (CNN, CNN-BiLSTM), that demonstrates how meticulous ensemble designs can provide high accuracy with reduced computational expenditure, that renders SER models more relevant to use in resource-limited real-world scenarios.

AUTHORS
First Author *–Ritu Vijay Bhalerao, M.Sc. Data Science, and Dr. D.Y Patil Arts, Commerce, Science College, Pimpri*
Second Author *–Kaveri Santosh Ahire, M.Sc. Data Science, and Dr. D.Y Patil Arts, Commerce, Science College, Pimpri*