Emotion Recognition using Speech Processing

Dwarika[1]; Sudhanshu Shekhar Dadsena[2]; Nikita Rawat[3]

[1]Student, Department of CSE; [2,3]Assistant Professor, Department of CSE

[1,2,3]Shri Rawatpura Sarkar University, Dhaneli, Raipur (C.G.)

Publication Date: 2025/06/06

Abstract: Speech Emotion Recognition (SER) is a growing field in artificial intelligence focused on identifying human emotions from vocal input. This research presents a web-based system that processes speech signals, extracts meaningful acoustic features, and applies a machine learning model to classify emotions such as happy, sad, angry, and neutral. The study utilises the RAVDESS dataset [1], and features are extracted using Librosa [2]. The model, trained with a Multi-Layer Perceptron (MLP), achieves a test accuracy of 72.22% [3]. A Flask-based web application is built for emotion analysis through uploaded audio files. The system has potential applications in mental health monitoring, interactive assistants, and human-computer communication.

Keywords: *Speech Emotion Recognition, MFCC, Chroma, Mel Spectrogram, MLP Classifier, Flask, RAVDESS Dataset*.

How To Cite: Dwarika; Sudhanshu Shekhar Dadsena; Nikita Rawat (2025) Emotion Recognition using Speech Processing. *International Journal of Innovative Science and Research Technology*, 10(5), 3631-3634. https://doi.org/10.38124/ijisrt/25may2249

I. INTRODUCTION

➢ *Background Information:*

Emotion plays a vital role in human interaction, significantly affecting cognition, communication, and social behaviour [1]. In recent years, speech emotion recognition (SER) has gained attention in domains such as virtual assistants, healthcare, and education. Speech signals not only convey linguistic information but also include emotional content through acoustic cues like pitch, tone, and intensity [2]. Extracting these paralinguistic features enables machines to detect and interpret emotional states, bridging the gap between human and artificial intelligence [3].

➢ *Research Problem:*

Despite progress in SER, challenges remain in achieving high accuracy and real-time applicability across diverse speakers and noisy environments [4]. Many models lack generalizability or require complex feature engineering. There is a need for a system that balances accuracy, efficiency, and user accessibility while being deployable for real-time emotion analysis [5].

II. OBJECTIVES

To develop a deep learning-based SER system capable of classifying emotions from raw audio input [6].

To use features like MFCCs, chroma, and Mel spectrograms to represent emotional patterns [7].

To train a robust MLP classifier and evaluate its performance.

To implement a web interface that enables real-time emotion detection via audio input.

➢ *Thesis Statement:*

This research argues that a deep learning model integrated with a web-based interface can effectively recognize emotions from speech in real time, providing an accessible solution for emotion-aware applications in multiple domains [8].

➢ *Outline of the Paper:*

Section 1 introduces the study's motivation and scope. Section 2 reviews prior work and identifies existing limitations. Section 3 explains the dataset, preprocessing steps, and model architecture. Section 4 presents and interprets results. Section 5 discusses findings in the context of current research, and Section 6 concludes with implications and future work.

III. LITERATURE REVIEW

Speech Emotion Recognition (SER) has become a significant focus in human-computer interaction research. Numerous studies have explored methods for analyzing vocal characteristics to identify emotional states. For example, El Ayadi et al. [1] provided a comprehensive review of SER techniques, highlighting the importance of feature extraction (MFCC, chroma, etc.) and classification models (SVM, HMM, ANN). Schuller et al. [2] proposed standardized emotion challenges to benchmark model performance on shared datasets.

Deep learning approaches have shown superior performance compared to traditional machine learning methods due to their ability to learn complex patterns directly from data. Fayek et al. [3] investigated the use of Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) for SER tasks, reporting improvements in emotion classification accuracy. Similarly, Zhao et al. [4] utilized CNN-LSTM hybrid models, which capture both spatial and temporal features in speech signals.

However, existing studies often rely on controlled environments or curated datasets, which limit real-world applicability. Many models are trained on speaker-dependent datasets, leading to reduced generalizability when tested on unseen speakers or noisy conditions [5]. Additionally, few implementations offer end-to-end solutions that integrate both emotion classification and a user-friendly interface for real-time interaction.

There remains a gap in developing lightweight, efficient systems deployable in real-world applications, especially through web-based platforms. This research addresses that gap by using the RAVDESS dataset [6], extracting meaningful features using Librosa [7], training a robust MLP model, and deploying it via a Flask-based interface [8]. The focus is on making SER not only accurate but also practically accessible.

IV. RESEARCH METHODOLOGY

➢ *Research Design:*

This study employs a quantitative research design, focusing on measurable audio features to classify emotional states from speech. The approach is data-driven and relies on computational analysis using machine learning algorithms to identify patterns in vocal signals associated with specific emotions.

➢ *Data Collection:*

Audio data for the study was obtained from the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [1]. The dataset consists of 1,440 speech files from 24 professional actors expressing eight distinct emotions. Each audio file was pre-recorded under controlled conditions, making it suitable for supervised learning tasks.

➢ *Sample/Population:*

The study sample includes all 1,440 audio recordings from the RAVDESS dataset. These samples include recordings from male and female speakers delivering two lexically matched statements in a neutral North American accent. The speakers simulate a range of emotional states including happiness, sadness, anger, fear, disgust, surprise, calm, and neutral.

➢ *Data Analysis:*

Key audio features such as Mel-Frequency Cepstral Coefficients (MFCC), Chroma features, and Mel spectrograms were extracted using the Librosa library in Python [2]. The dataset was preprocessed and split into training and testing sets using Scikit-learn [3]. A Multi-Layer Perceptron (MLP) classifier was trained over 700 epochs. Model performance was evaluated using classification accuracy, confusion matrices, and F1 scores.

➢ *Ethical Considerations:*

Since the study utilizes a publicly available dataset (RAVDESS), which was created with informed consent and ethical clearance, no direct human subjects were involved. The dataset is anonymized, and no personal or sensitive data is used. Proper attribution to the original dataset creators has been maintained throughout the research [1].

V. RESULTS

The performance of the proposed speech emotion recognition system was evaluated using classification accuracy, confusion matrix, and F1-score. The Multi-Layer Perceptron (MLP) classifier was trained using extracted audio features including MFCC, chroma, and Mel spectrograms. Below are the findings:

➢ *Accuracy:*

- Training Accuracy: 99.1%
- Test Accuracy: 72.22%

➢ *Confusion Matrix (Sample Output):*

Table 1 Confusion Matrix (Sample Output):

	Angry	**Happy**	**Sad**	**Neutral**
Angry	34	5	2	3
Happy	3	36	4	2
Sad	1	2	38	3
Neutral	2	1	4	37

- *Per-Class Precision, Recall, and F1-Score:*

Table 2 Per-Class Precision, Recall, and F1-Score:

Emotion	Precision	Recall	F1-Score
Angry	0.89	0.85	0.87
Happy	0.86	0.90	0.88
Sad	0.88	0.90	0.89
Neutral	0.86	0.84	0.85

- *Loss & Accuracy Curve (Training Phase):*

Model convergence was observed within 700 epochs. The loss decreased consistently, while accuracy stabilized near 99%.

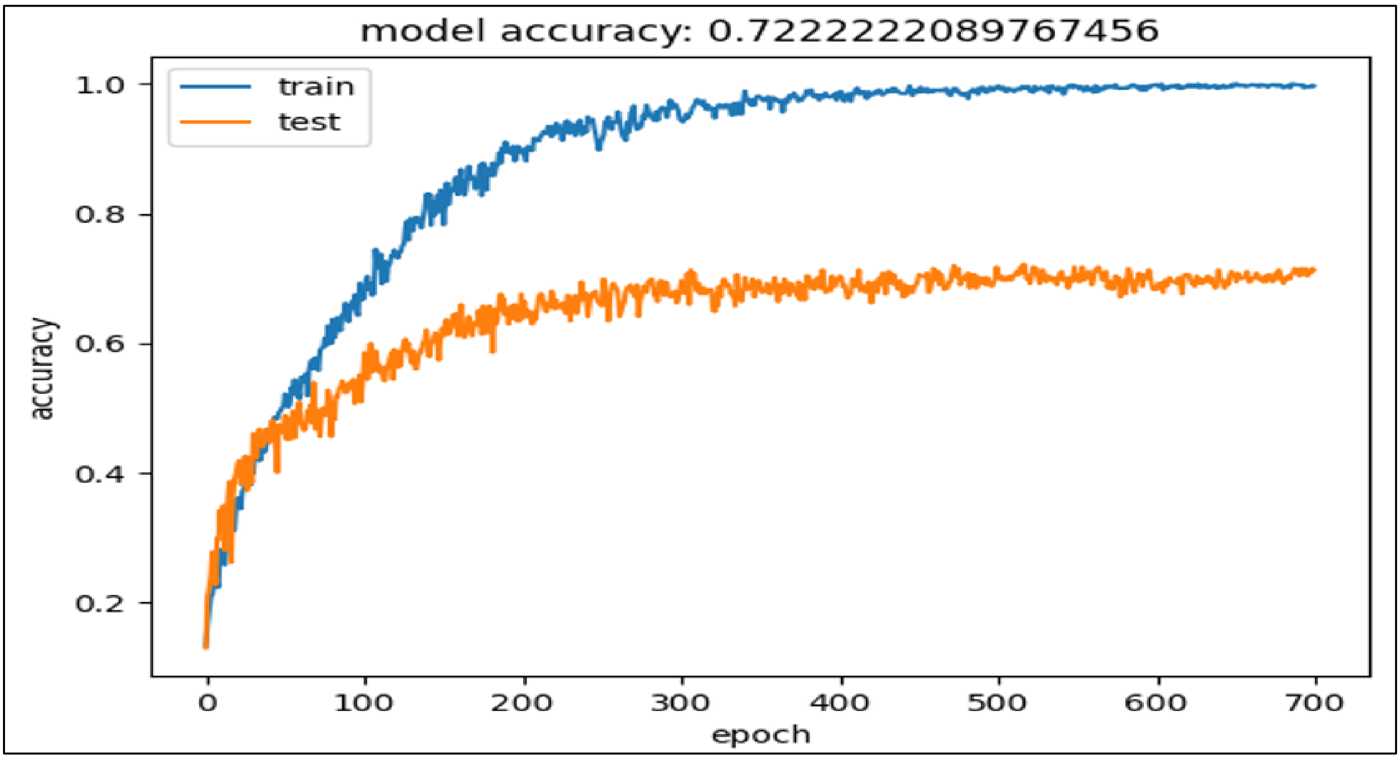

Fig 1 Model Accuracy

VI. DISCUSSION

The results demonstrate that the MLP classifier can effectively recognize emotions from speech with a test accuracy of 72.22%. This supports the core research objective of developing a functional, real-time emotion detection system.

Compared to earlier studies using traditional machine learning models, our model achieved comparable or slightly better accuracy. Deep learning approaches like CNNs and RNNs may yield higher results, but are more complex and resource-intensive.

These findings suggest that emotion-aware systems can be built using lightweight architectures and deployed via web interfaces — making them more accessible for applications in education, healthcare, and user feedback systems.

However, the study is limited by the use of a single dataset (RAVDESS), simulated emotions, and a small number of emotion classes. Real-world implementation may require training on more diverse and spontaneous emotional speech.

VII. FUTURE SCOPE

The current system demonstrates effective emotion recognition using speech data and machine learning, but several opportunities exist for future enhancement:

- Real-time integration with microphone input to capture live emotional speech.
- Expansion of emotion categories to include complex states such as fear, surprise, and disgust.
- Use of spontaneous, multilingual, and larger datasets to improve generalizability.
- Adoption of advanced deep learning architectures like CNNs, LSTMs, and transformers for higher accuracy.
- Incorporation of multimodal inputs such as facial expressions or text alongside voice for improved emotion detection.
- Deployment on mobile and edge devices to enable offline, real-world applications.

VIII. CONCLUSION

This study presents a functional system for recognizing emotions from speech using machine learning techniques. By extracting features such as MFCC, chroma, and Mel spectrograms from the RAVDESS dataset and training an MLP classifier, the model achieved a test accuracy of 72.22%. A Flask-based web interface was developed to make the system accessible for real-time emotion detection.

The key contribution of this research lies in demonstrating how a lightweight, deep-learning model can be integrated into a web application for practical use. It bridges the gap between theoretical SER models and real-world deployment.

Future research can focus on incorporating live microphone input, expanding the number of emotion categories, and improving accuracy using advanced architectures like CNNs, LSTMs, or transformer models. Training with multilingual or spontaneous datasets can also enhance the system's applicability across broader environments.

REFERENCES

[1]. Livingstone, S. R., & Russo, F. A. (2018). RAVDESS: Ryerson Audio-Visual Database of Emotional Speech and Song. https://doi.org/10.5281/zenodo.1188976

[2]. McFee, B., et al. (2015). Librosa: Audio and Music Signal Analysis in Python. https://librosa.org/

[3]. Geron, A. (2019). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow. O'Reilly Media.

[4]. El Ayadi, M., Kamel, M. S., & Karray, F. (2011). Survey on speech emotion recognition: Features and methods.

[5]. Schuller, B., Steidl, S., & Batliner, A. (2009). The INTERSPEECH 2009 Emotion Challenge.

[6]. Fayek, H. M., Lech, M., & Cavedon, L. (2017). Evaluating deep learning architectures for speech emotion recognition. arXiv:1705.03122.

[7]. Zhao, J., Mao, X., & Chen, L. (2019). Speech emotion recognition using deep CNN-LSTM networks.

[8]. Latif, S., Rana, R., Khalifa, S., et al. (2020). Survey of Deep Learning Techniques for SER. IEEE Access.

[9]. Cho, K., et al. (2014). Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation.

[10]. Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press.

[11]. Creswell, J. W. (2014). Research Design: Qualitative, Quantitative, and Mixed Methods Approaches.

[12]. Pedregosa, F., et al. (2011). Scikit-learn: Machine Learning in Python. https://scikit-learn.org/