IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 3, March 2026

DOI: 10.17148/IARJSET.2026.13375

# SPEECH EMOTION RECOGNITION USING MACHINE LEARNING

**MADHAN E[1], Dr. A. ADHISELVAM[2]**

B.Sc. Information Technology, Dr. N.G.P. Arts and Science College, Coimbatore[1]

Professor and Head, Department Dr. N.G.P. Arts and Science College, Coimbatore[2]

**Abstract:** Speech Emotion Recognition (SER) is an important area of research in human–computer interaction that aims to identify human emotions from speech signals. Accurate detection of emotions such as happiness, sadness, anger, fear, and neutrality can significantly enhance applications in virtual assistants, mental health monitoring, and customer service systems. Traditional emotion recognition systems relied on handcrafted acoustic features and conventional machine learning techniques, which often struggled to capture complex patterns in speech data.

In this project, a deep learning–based approach is proposed to automatically recognize emotions from speech signals. The system processes audio inputs by extracting relevant acoustic features such as Mel-Frequency Cepstral Coefficients (MFCCs), pitch, and energy. These features are then used to train a deep learning model capable of learning emotional patterns from speech data. The proposed model improves classification accuracy by effectively capturing temporal and spectral characteristics of speech signals.

The developed system is evaluated using labeled speech emotion datasets and demonstrates promising performance in recognizing multiple emotional states. The results show that deep learning models can significantly enhance emotion recognition accuracy compared to traditional approaches. This work highlights the potential of speech emotion recognition systems in building more natural and emotionally aware human–machine interactions.

**Keywords:** Artificial Intelligence, Machine Learning, emotion Detection,

## I. INTRODUCTION

Speech is one of the most natural and effective ways of communication among humans. Apart from conveying linguistic information, speech also carries emotional cues that reflect the speaker's feelings and mental state. Recognizing emotions from speech plays a significant role in improving human–computer interaction by enabling machines to understand and respond to human emotions more effectively. Speech Emotion Recognition (SER) is a field of research that focuses on automatically identifying human emotions from speech signals using computational methods.

In recent years, SER has gained increasing attention due to its wide range of applications in areas such as virtual assistants, call center analysis, healthcare monitoring, and intelligent tutoring systems. By detecting emotions such as happiness, sadness, anger, fear, and neutrality, SER systems can enhance the responsiveness and adaptability of interactive technologies. However, accurately identifying emotions from speech remains a challenging task because emotional expressions vary across individuals, languages, and speaking styles.

Traditional approaches for speech emotion recognition relied mainly on handcrafted acoustic features such as pitch, energy, and Mel-Frequency Cepstral Coefficients (MFCCs), combined with machine learning algorithms like Support Vector Machines (SVM) and Hidden Markov Models (HMM). Although these methods provided reasonable results, they often struggled to capture complex patterns present in speech signals.

## II. LITERATURE REVIEW

Earlier research on speech emotion recognition mainly relied on traditional signal processing and machine learning techniques. **Lawrence Rabiner (1989)** introduced methods based on **Hidden Markov Model** for analyzing speech patterns. These methods utilized handcrafted acoustic features such as pitch, energy, and spectral properties to classify emotions. Although these approaches provided a foundation for speech analysis, they were limited in capturing the complex and dynamic nature of emotional expressions in speech.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 3, March 2026

DOI: 10.17148/IARJSET.2026.13375

With the development of machine learning techniques, **Vladimir Vapnik (1995)** proposed the **Support Vector Machine**, which became widely used for emotion classification tasks. Many researchers applied SVM models with acoustic features such as **Mel-Frequency Cepstral Coefficients (MFCCs)** to detect emotions in speech signals. While these models improved classification performance compared to earlier statistical approaches, they still depended heavily on manually engineered features and struggled with large and complex datasets.

With the rise of deep learning, researchers began applying neural networks for emotion recognition. **Alex Krizhevsky et al. (2012)** demonstrated the effectiveness of **Convolutional Neural Network (CNN)** in automatically extracting meaningful patterns from data. Inspired by this success, CNN-based models were later applied to spectrogram representations of speech signals to recognize emotional states more accurately. These models significantly improved performance by learning hierarchical features directly from the input data.

To further enhance speech modeling, **Sepp Hochreiter and **Jürgen Schmidhuber (1997)** introduced **Long Short-Term Memory (LSTM)** networks, which are capable of capturing long-term dependencies in sequential data such as speech. LSTM-based models have been widely used in SER systems because they effectively model the temporal dynamics of speech signals. However, these models can be computationally intensive and require large datasets for training.

More recently, hybrid deep learning architectures combining CNN and LSTM have been proposed to improve both spatial and temporal feature extraction from speech data. These models achieve higher accuracy in recognizing emotions by analyzing both spectral patterns and time-based variations in speech signals. Despite these improvements, challenges such as speaker variability, noise, and limited labeled datasets still affect the performance of speech emotion recognition systems.

These studies demonstrate the growing effectiveness of deep learning techniques for speech emotion recognition. However, there is still a need for systems that can achieve high accuracy while maintaining computational efficiency. The proposed system aims to address these challenges by implementing an efficient deep learning–based approach for recognizing emotions from speech signals, improving the overall performance and reliability of SER applications.

## III. PROJECT OVERVIEW

Speech Emotion Recognition (SER) is a technology that enables computers to identify human emotions from speech signals. Human speech not only conveys linguistic information but also reflects emotional states such as happiness, sadness, anger, fear, and neutrality. Recognizing these emotions can significantly enhance human–computer interaction by allowing machines to respond more naturally and intelligently to users.

The primary objective of this project is to develop a system that can automatically detect and classify emotions from speech using deep learning techniques. The system processes audio input, extracts important acoustic features such as **Mel-Frequency Cepstral Coefficients (MFCCs)**, pitch, and energy, and then uses these features to train a deep learning model capable of identifying emotional patterns in speech.

.

**The project consists of the following main components:**

1. Speech Input Module

This component is responsible for collecting speech data from the user or from a dataset. The audio input serves as the primary source for detecting emotional patterns in speech.

2. Preprocessing Module

In this stage, the raw speech signals are cleaned and prepared for analysis. Noise removal, normalization, and signal enhancement techniques are applied to improve the quality of the audio data.

3. Feature Extraction Module

This component extracts meaningful features from the speech signals that represent emotional characteristics. Commonly used features include Mel-Frequency Cepstral Coefficients (MFCCs), pitch, and energy, which help capture important acoustic information from the speech signal.

4. Deep Learning Model Module

The extracted features are fed into a deep learning model such as a Convolutional Neural Network (CNN) or Long Short-Term Memory (LSTM). These models learn complex patterns in speech data and classify different emotional states.

5. Emotion Classification ModuleThe trained model analyzes the processed speech input and predicts the emotion expressed by the speaker. The system classifies the speech into categories such as happiness, sadness, anger, fear, or neutral emotion.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 3, March 2026

DOI: 10.17148/IARJSET.2026.13375

6. Output Display Module

The final component presents the detected emotion to the user. The results may be displayed as emotion labels, probability scores, or graphical representations for better understanding.

.

## IV. ALGORITHM DESCRIPTION

- InputSpeechSignal The system accepts an audio file or real-time speech input from the user. This speech signal contains both linguistic and emotional information.
- Audio Preprocessing The input speech signal is preprocessed to improve its quality. Noise removal, normalization, and segmentation are performed to make the signal suitable for analysis.
- FeatureExtraction Important acoustic features are extracted from the speech signal. One of the most commonly used features is Mel-Frequency Cepstral Coefficients (MFCCs), which represent the spectral characteristics of speech. Additional features such as pitch and energy may also be extracted.
- FeatureRepresentationThe extracted features are converted into a structured numerical format that can be used as input for a deep learning model.
- ModelTrainingThe feature dataset is used to train a deep learning model such as a Convolutional Neural Network (CNN) or Long Short-Term Memory (LSTM) network. The model learns patterns associated with different emotions from the training data.
- EmotionPredictionWhen a new speech sample is provided, the system extracts features from the input and feeds them into the trained model. The model analyzes the patterns and predicts the corresponding emotional category.
- OutputGenerationFinally, the system displays the predicted emotion (such as happiness, sadness, anger, or neutral) as the output.

## V. SYSTEM ANALYSIS

System analysis is an important phase in the development of the Speech Emotion Recognition (SER) system. It involves studying the existing systems, identifying their limitations, and defining the requirements for the proposed system. The purpose of system analysis is to understand how speech data can be processed and analyzed effectively to detect human emotions

### 5.1 EXISTING SYSTEM

In the existing systems, speech emotion recognition was mainly performed using traditional machine learning techniques. These systems relied on manually designed acoustic features such as **Mel-Frequency Cepstral Coefficients (MFCCs)**, pitch, and energy, which were then classified using algorithms like **Support Vector Machine (SVM)** or **Hidden Markov Model (HMM)**. Although these methods provided basic emotion recognition capabilities, they often struggled to capture complex emotional patterns in speech signals. As a result, the accuracy of emotion classification was limited, especially in noisy environments or when dealing with large datasets.

**Limitations of Existing System**

- Traditional methods rely heavily on handcrafted features.
- Difficulty in capturing complex emotional patterns in speech signals.
- Lower accuracy in real-world environments with background noise.
- Limited ability to handle large and diverse speech datasets.
- Poor performance in recognizing subtle emotional variations

### 5.2 PROPOSED SYSTEM

The proposed system introduces a deep learning–based approach to improve emotion recognition performance. Instead of relying only on manual feature engineering, deep learning models can automatically learn meaningful representations from speech data. In this system, speech signals are processed, and important features such as **Mel-Frequency Cepstral Coefficients (MFCCs)** are extracted and fed into deep learning models like **Convolutional Neural Network (CNN)** or **Long Short-Term Memory (LSTM)**.

These models can effectively learn both spatial and temporal patterns in speech signals, resulting in improved emotion recognition accuracy. The proposed system provides a more reliable and scalable solution for detecting emotions from speech data.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 3, March 2026

DOI: 10.17148/IARJSET.2026.13375

**Advantages of Proposed System**

- Higher emotion recognition accuracy.
- Ability to automatically learn complex patterns in speech signals.
- Better performance with large datasets.
- Improved robustness to noise and variations in speech.
- Enhanced human–computer interaction through emotion-aware systems

## VI. SYSTEM OVERVIEW

The **Speech Emotion Recognition (SER) system using Machine Learning** is designed to automatically identify human emotions from speech signals. Human speech contains important emotional information such as tone, pitch, intensity, and rhythm. By analyzing these features, the system can classify emotions like **happy, sad, angry, fear, neutral, and surprise**.

The system works by capturing a speech input and converting it into a digital signal. The audio signal is then processed using various preprocessing techniques such as **noise reduction, normalization, and segmentation** to improve the quality of the data. After preprocessing, important audio features like **Mel Frequency Cepstral Coefficients (MFCC), pitch, energy, and spectral features** are extracted from the speech signal.

These extracted features are used as input for a **machine learning model** that has been trained on a dataset containing labeled emotional speech samples. The model learns patterns associated with different emotions during the training phase. Once the system receives a new speech input, it extracts the same features and feeds them into the trained model. The model then predicts the **most likely emotional state** expressed in the speech.

The system is implemented using **Python programming language**, along with machine learning libraries such as **Librosa, NumPy, Scikit-learn, and TensorFlow/Keras** for audio processing and model training. A simple **Flask-based interface** can be used to allow users to upload or record speech and receive emotion predictions

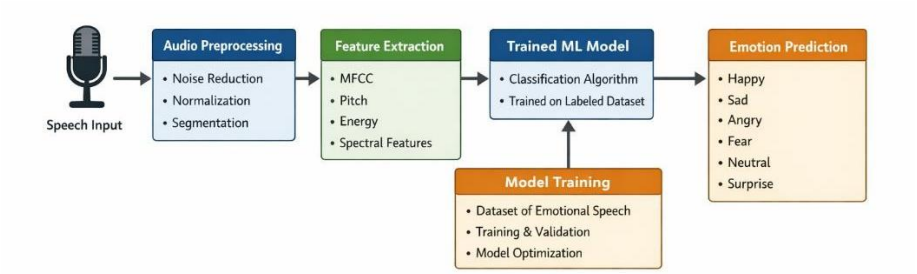


Fig:1 System Flow Diagram

## VII. MODULE DESCRIPTION

The Speech Emotion Recognition system is divided into several modules. Each module performs a specific function in processing the speech signal and predicting the emotion.

**1. Speech Input Module**

This module is responsible for collecting the speech signal from the user. The input speech can be obtained through a **microphone recording** or by uploading an **audio file**. The speech signal is converted into a digital format that can be processed by the system.

**2. Audio Preprocessing Module**

In this module, the raw speech signal is cleaned and prepared for analysis. The preprocessing techniques help improve the quality of the audio data and remove unwanted noise. Main tasks include:

- **Noise Reduction** – Removing background noise from the audio signal.
- **Normalization** – Adjusting the amplitude of the signal for consistent processing.

**3. Feature Extraction Module**

This module extracts important characteristics from the speech signal that represent emotional patterns. These features help the machine learning model understand variations in speech tone and energy.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 3, March 2026

DOI: 10.17148/IARJSET.2026.13375

Common extracted features include:

- **Mel Frequency Cepstral Coefficients (MFCC)**
- **Pitch**
- **Energy**
- **Spectral Features**

**4. Machine Learning Model Module**

This module uses a trained machine learning algorithm to classify emotions from the extracted features. The model is trained using a labeled dataset of emotional speech samples. The algorithm learns patterns associated with different emotions during the training phase.

**5. Emotion Classification Module**

In this module, the trained model analyzes the extracted features and predicts the emotional state of the speaker. The system classifies the emotion into categories such as:

- Happy
- Sad
- Angry
- Fear
- Neutral
- Surprise

**6. Output Display Module**

This module displays the predicted emotion to the user through the system interface. The result can be shown as **text output, graphical visualization, or emotion labels** depending on the application design

.

## VIII. RESULTS AND DISCUSSION

The Speech Emotion Recognition system was successfully implemented using machine learning techniques to identify emotions from speech signals. The system was trained using a dataset containing speech samples labeled with different emotional states.

During testing, the system was able to classify emotions such as **happy, sad, angry, fear, neutral, and surprise** from the given speech input. After preprocessing the audio signal and extracting important features such as **Mel Frequency Cepstral Coefficients (MFCC), pitch,**

.

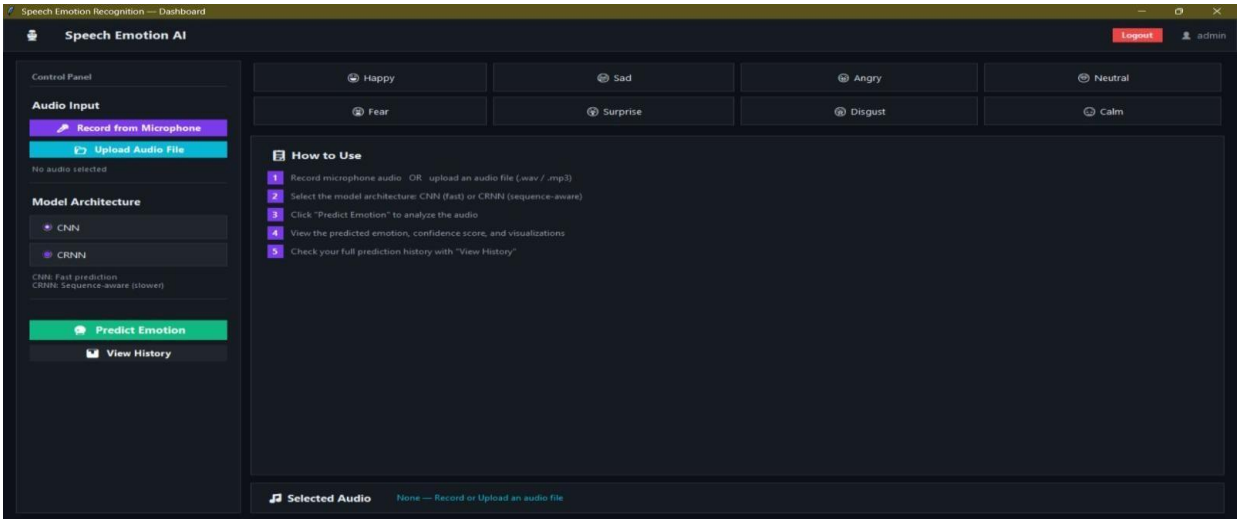


Fig:2 emotion recognition

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 3, March 2026

DOI: 10.17148/IARJSET.2026.13375


Fig:2.1audio selecting

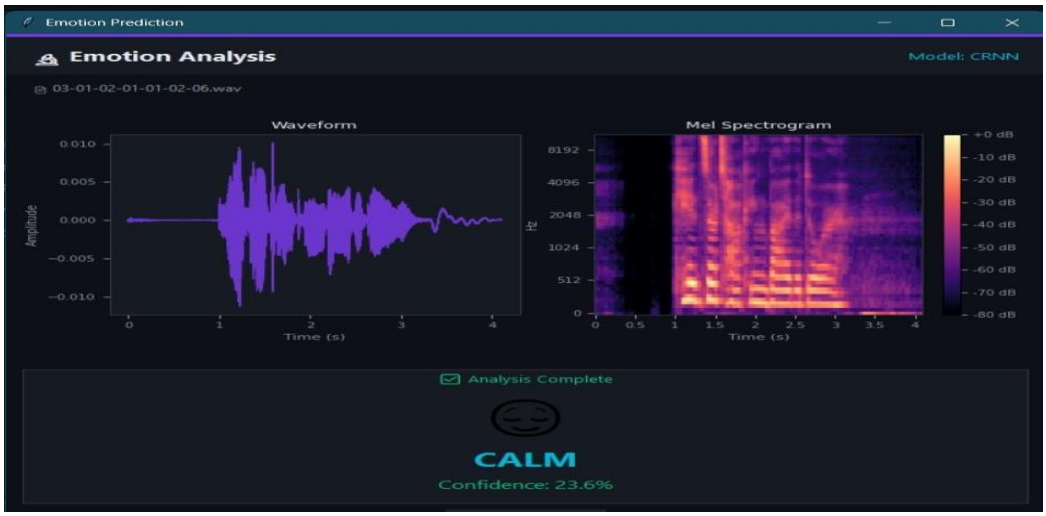


Fig:2.2 Analysis Dashboard

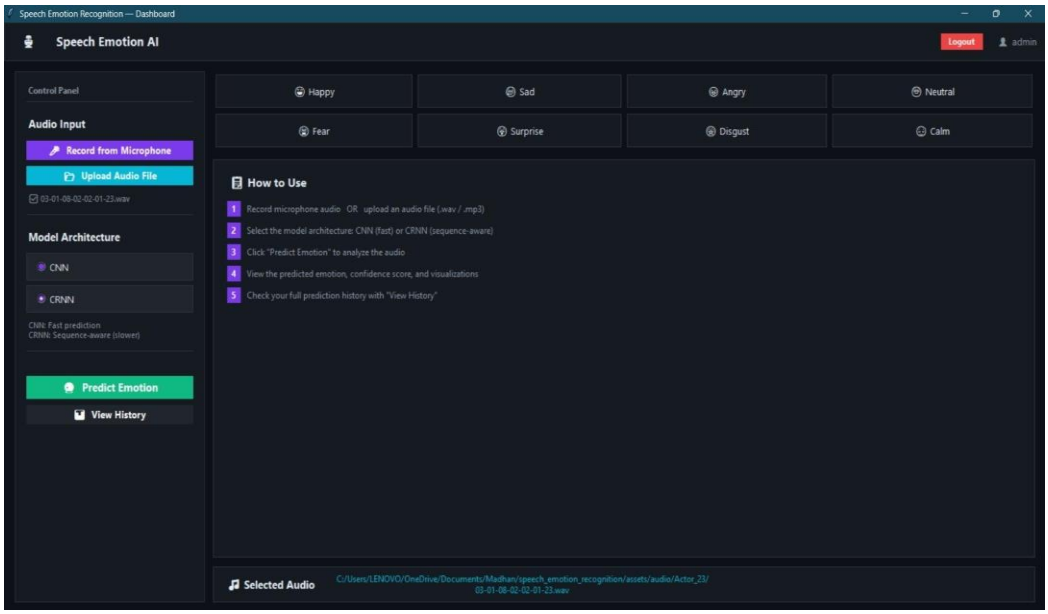


Fig:2.3 Analytics Dashboard

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 3, March 2026

DOI: 10.17148/IARJSET.2026.13375

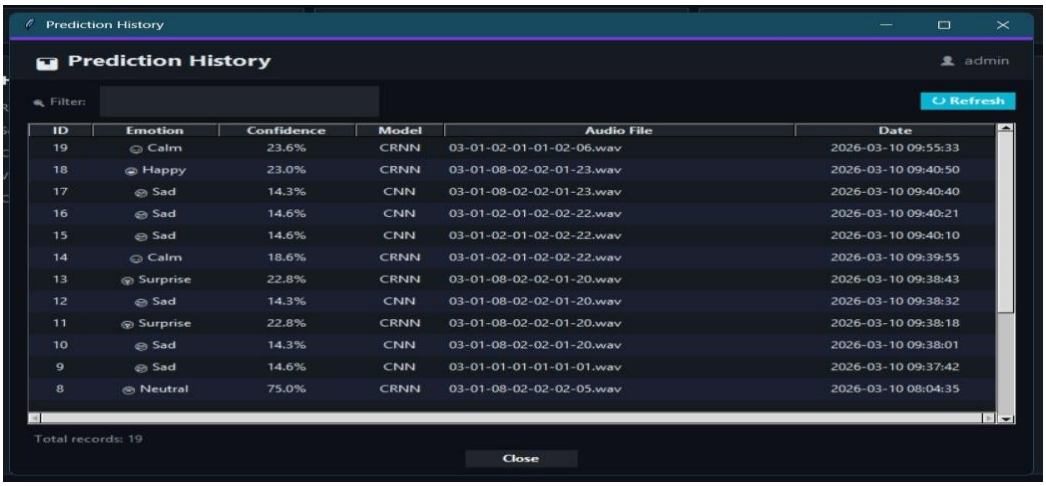


Fig:2.4 Prediction History

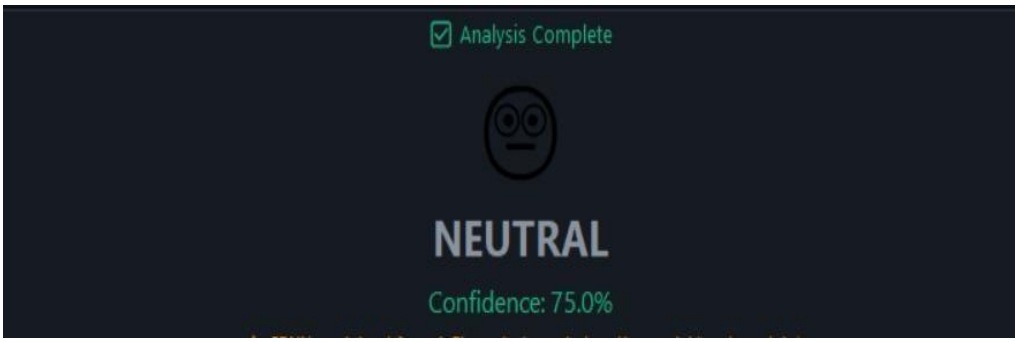


Fig:2.5 prediction result

## IX. CONCLUSION

In this project, a **Speech Emotion Recognition (SER) system using Machine Learning** was developed to identify human emotions from speech signals. The system processes speech input, performs audio preprocessing, extracts relevant features, and uses a trained machine learning model to classify emotions.

Important speech features such as **Mel Frequency Cepstral Coefficients (MFCC), pitch, and energy** were extracted from the audio signals to represent emotional characteristics. These features were then used by the machine learning model to recognize emotions such as **happy, sad, angry, fear, neutral, and surprise**

## REFERENCES

[1]. L. R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, Feb. 1989.

[2]. V. Vapnik, *The Nature of Statistical Learning Theory*. New York, NY, USA: Springer, 1995.

[3]. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in *Proceedings of the 25th International Conference on Neural Information Processing Systems (NIPS)*, 2012, pp. 1097–1105.

[4]. S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[5]. B. Schuller, S. Steidl, and A. Batliner, "The INTERSPEECH 2009 Emotion Challenge," in *Proceedings of Interspeech*, 2009, pp. 312–315.

[6]. Z. Huang, M. Dong, Q. Mao, and Y. Zhan, "Speech Emotion Recognition Using CNN and LSTM," *IEEE Transactions on Affective Computing*, vol. 10, no. 3, pp. 1–10, 2019.