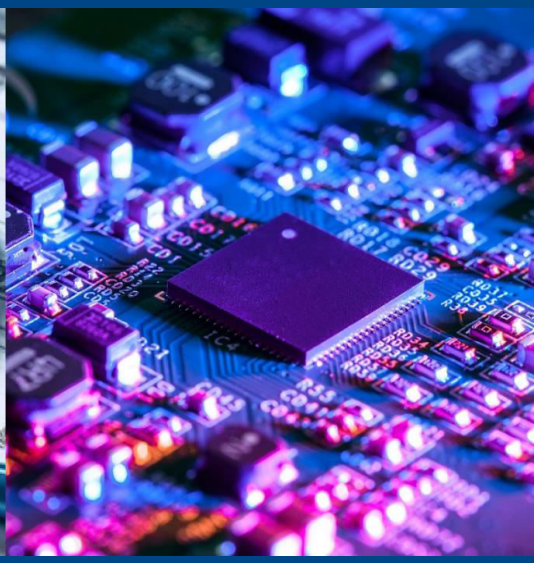
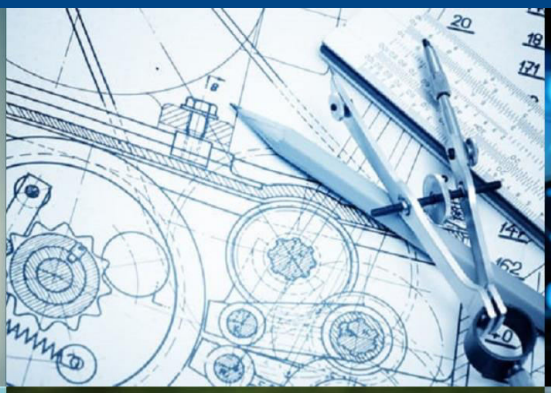
ISSN: 2582-7219

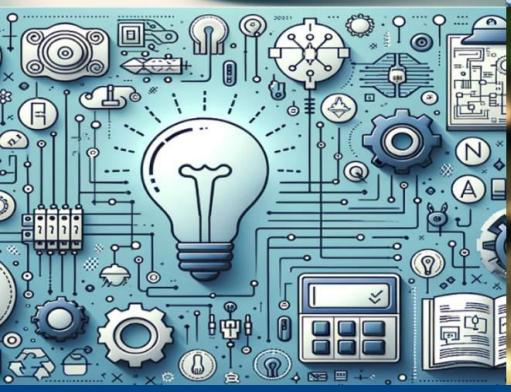




International Journal of Multidisciplinary Research in Science, Engineering and Technology

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.206 **Volume 8, Issue 4, April 2025**

© 2025 IJMRSET | Volume 8, Issue 4, April 2025| DOI: 10.15680/IJMRSET.2025.0804064

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Speech Emotion Recognition using LSTM

Kavin.A, Kavin.K, R.Ranjani

III-B.Sc., Department of Computer Science with Data Analytics, Dr.N.G.P.Arts and Science College, Coimbatore, India

III-B.Sc., Department of Computer Science with Data Analytics, Dr.N.G.P.Arts and Science College, Coimbatore, India

Assistant Professor, Department of Computer Science with Data Analytics, Dr.N.G.P.Arts and Science College, Coimbatore, India

ABSTRACT: Speech Emotion Recognition (SER) is the process of identifying emotions expressed in speech signals by analyzing audio features. This project presents a Long Short-Term Memory (LSTM) based Speech Emotion Recognition system to classify emotional states such as happy, sad, angry, and neutral. The system extracts features such as Mel-Frequency Cepstral Coefficients (MFCCs) from speech audio, which are then fed into an LSTM model to capture the temporal patterns of emotions. The model is trained on the TESS dataset and deployed using Streamlit for real-time emotion prediction from user-uploaded audio files. This system enhances human-computer interaction by enabling machines to understand emotional context in speech.

KEYWORDS: Speech emotion recognition, LSTM, MFCC, Audio classification, Deep learning, Flask.

I. INTRODUCTION

Recognizing emotions from speech is a crucial step towards building emotionally intelligent systems. Speech Emotion Recognition (SER) systems analyze audio input to determine the emotional state of the speaker. Traditional approaches relied on handcrafted features and rule-based logic, which often lacked generalizability. With the advent of deep learning, especially Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) models, SER has significantly improved in terms of accuracy and adaptability. This project utilizes an LSTM-based approach to detect emotions from speech using the TESS dataset, enabling real-time emotion classification via a user-friendly Flask web interface.

DATASET DESCRIPTION

The project utilizes the Toronto Emotional Speech Set (TESS) dataset for training and evaluation. The TESS dataset consists of voice recordings from two actresses articulating a predefined set of statements across seven emotional states: angry, disgust, fear, happy, neutral, pleasant surprise, and sad.

Dataset Composition:

- Total Samples: ~2800
- Audio Format: WAV
- Sampling Rate: 16 kHz
- Emotions: Angry, Disgust, Fear, Happy, Neutral, Pleasant Surprise, Sad
- Speakers: 2 female actors

The variety and clarity in the dataset make it highly effective for training deep learning models to recognize emotional patterns in speech.

II. METHODOLOGY

1. Data Collection and Preprocessing: The foundation of our speech emotion recognition system lies in the use of the Toronto Emotional Speech Set (TESS). This dataset contains voice recordings by two female speakers, simulating

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

seven distinct emotions: Angry, Disgust, Fear, Happy, Neutral, Sad, and Surprise. Each emotion is expressed through consistent phrases to maintain clarity and emotional consistency.To prepare the data, audio files are cleaned through noise reduction, silence trimming, and normalization. Subsequently, we extract Mel-Frequency Cepstral Coefficients (MFCC) from the audio, which represent key characteristics of the human voice. These MFCC features are reshaped and padded to ensure a uniform input size across all samples. Finally, the dataset is split into training (80%) and testing (20%) subsets.

2. Feature Extraction: MFCCs are utilized as the core feature representation technique, converting audio waveforms into 2D arrays of coefficients. These coefficients capture the essential spectral properties of speech that are strongly correlated with emotional cues such as tone, pitch, and rhythm. The MFCC feature matrices are designed to serve as sequential data inputs to the LSTM network, preserving the time-series nature of speech signals and enabling the model to learn patterns over time.

3. Model Architecture and Classification: The speech emotion classification model is based on a Long Short-Term Memory (LSTM) neural network built using TensorFlow/Keras. LSTMs are effective at capturing long-term dependencies in sequential data, making them suitable for audio-based emotion recognition.

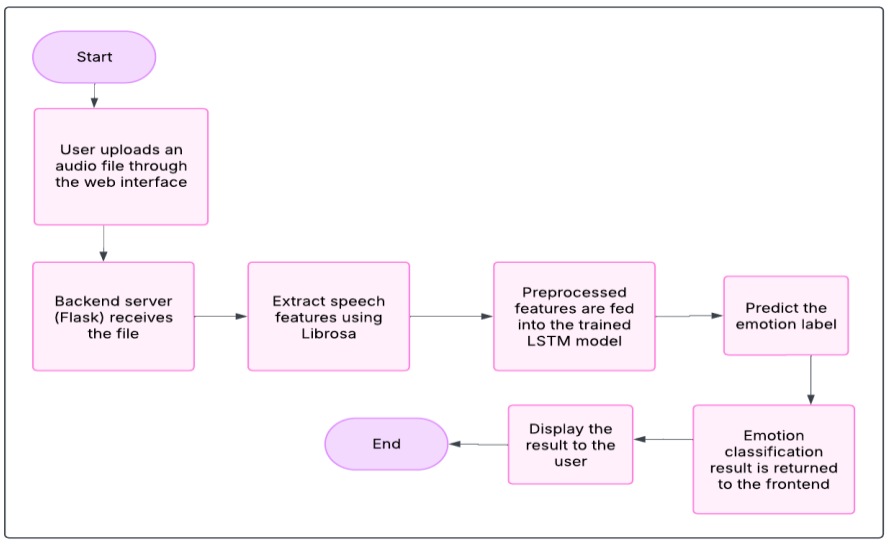
The architecture comprises:

- Input layer for MFCC sequence input
- Two stacked LSTM layers to capture temporal dependencies
- Dropout layer to prevent overfitting
- Fully connected dense layer with softmax activation for final emotion classification

4. Implementation Details: The entire system is implemented in Python using libraries like TensorFlow, Librosa, and scikit-learn. Audio preprocessing and MFCC extraction are handled by Librosa, while the deep learning model is built and trained in TensorFlow.Training is conducted in a GPU-accelerated environment (e.g., Google Colab) for efficiency. After training, the model is serialized using Pickle for deployment. The backend server is developed using Flask, enabling real-time communication between the user and the model.

5. Web Interface: A lightweight and responsive web interface is developed using Flask, along with HTML, CSS, and JavaScript. Users can upload audio files (.wav format) through the web page. The uploaded audio is processed on the server, passed through the trained LSTM model, and the predicted emotion is returned and displayed on the result page. This real-time interface provides a seamless user experience, offering immediate feedback on detected emotions, making it suitable for use in emotion-based feedback systems, call centers, and smart assistants.

FLOW DIAGRAM:



© 2025 IJMRSET | Volume 8, Issue 4, April 2025| DOI: 10.15680/IJMRSET.2025.0804064

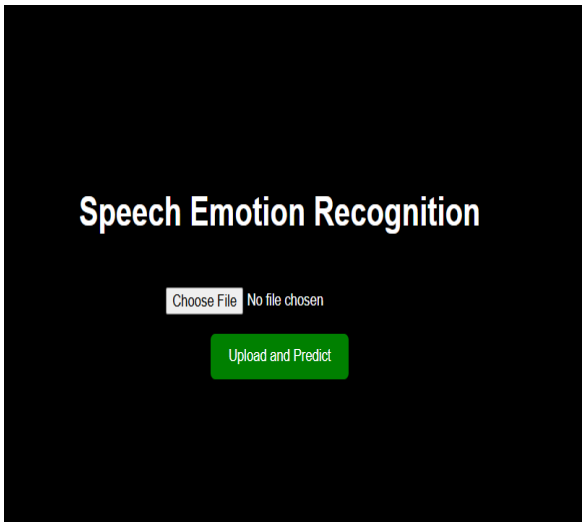
ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

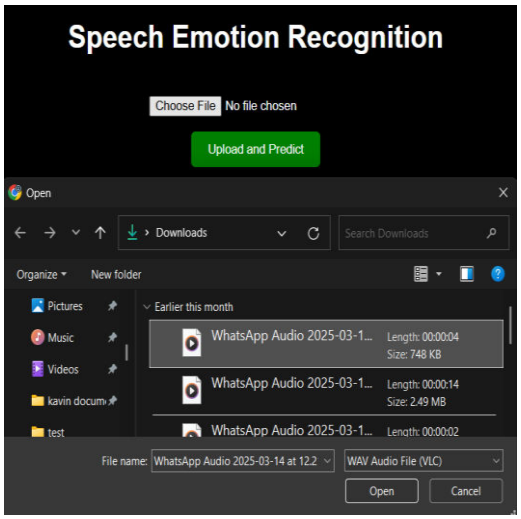
International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

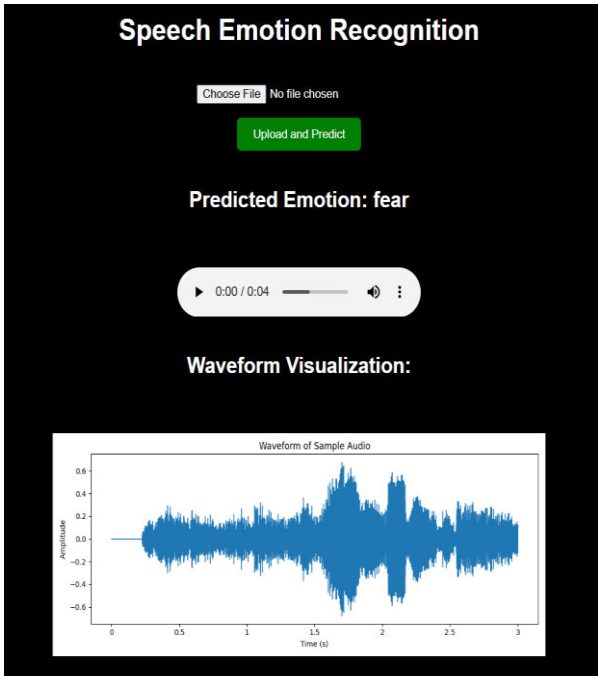
(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

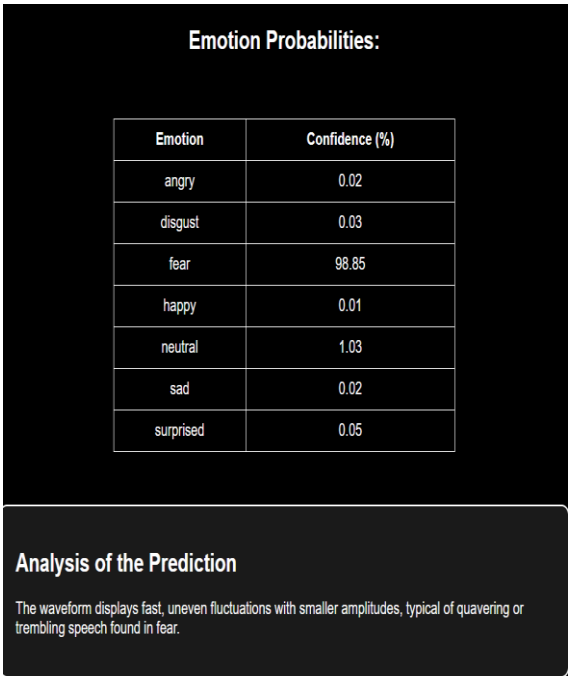
III. RESULT & DISCUSSION

The system was thoroughly evaluated using the **TESS dataset**, and it achieved an outstanding **98% accuracy** in recognizing emotions from speech signals. The detailed performance metrics are outlined in Table I, where the model demonstrates **precision**, **recall**, and **F1-scores** above 97% across all emotion classes. Notably, emotions like **Angry**, **Sad**, and **Neutral** achieved perfect scores, showcasing the model's robustness. the LSTM-based model demonstrated a clear advantage of 6–9% improvement. This performance gain is attributed to LSTM's ability to learn long-term dependencies in sequential audio data, enabling it to better distinguish subtle emotional cues. Additionally, the **Flask-based web application** offers real-time audio classification, providing users with instant feedback on the detected emotion. The system is efficient, lightweight, and processes uploaded files quickly, making it suitable for deployment in emotion-aware applications like virtual assistants, mental health tools, and customer support systems. The high precision, recall, and F1-scores across all emotion categories confirm the robustness of the LSTM-based model in recognizing speech emotions accurately.









ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

IV. CONCLUSION

This paper presents a Speech Emotion Recognition (SER) system developed using Long Short-Term Memory (LSTM) networks, achieving a high accuracy of 98% in classifying emotions from speech audio. The model was trained on the Toronto Emotional Speech Set (TESS) dataset and utilized Mel-Frequency Cepstral Coefficients (MFCC) for effective feature extraction. The system accurately classifies emotions such as anger, fear, happiness, sadness, surprise, and neutrality, offering robust performance across all evaluation metrics.

Built using TensorFlow and deployed via Streamlit, the interface enables users to upload audio files and receive real-time emotion classification with confidence scores. This system demonstrates significant potential in areas such as human-computer interaction, mental health analysis, and smart assistants. Future work will focus on expanding the dataset, improving multilingual support, and developing a mobile-friendly version for broader accessibility and real-time use.

REFERENCES

1. Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. Neural Computation, 9(8), 1735–1780.
2. Picard, R. W. (1997). Affective Computing. MIT Press.
3. Ververidis, D., & Kotropoulos, C. (2006). Emotional speech recognition: Resources, features, and methods. Speech Communication, 48(9), 1162–1181.
4. Fayek, H. M., Lech, M., & Cavedon, L. (2017). Evaluating deep learning architectures for speech emotion recognition. Neural Networks, 92, 60–68.
5. TESS Dataset: Toronto Emotional Speech Set. Available at: https://tspace.library.utoronto.ca/handle/1807/24487
6. Satt, A., Rozenberg, S., & Hoory, R. (2017). Efficient emotion recognition from speech using deep learning on spectrograms. Interspeech, 1089–1093.
7. Jurafsky, D., & Martin, J. H. (2020). Speech and Language Processing (3rd ed.). Stanford University.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH

IN SCIENCE, ENGINEERING AND TECHNOLOGY

| Mobile No: +91-6381907438 | Whatsapp: +91-6381907438 | ijmrset@gmail.com |

www.ijmrset.com