**Project Progress Report on**



**Speech Emotion Recognition (Classification) in real- time using Deep LSTM layers**



### Submitted in partial fulfilment of the requirement for the award of the degree of

**BACHELOR OF TECHNOLOGY IN**

### COMPUTER SCIENCE & ENGINEERING

**Submitted by:**

|  |  |
| --- | --- |
| **Ishika Gupta** | **2016787** |
| **Rohan Pandey** | **2016967** |
| **Yuganshu Joshi** | **2017147** |

***Under the Guidance of***

**Dr. Ashwini Kumar Singh**

### Associate Professor

**Project Team ID: MP23CSE152**



**Department of Computer Science and Engineering Graphic Era (Deemed to be University)**

**Dehradun, Uttarakhand April-2024**



**CANDIDATE’S DECLARATION**

We hereby certify that the work which is being presented in the Project Progress Report entitled **“Speech Emotion Recognition (Classification) in real-time using Deep LSTM layers”** in partial fulfillment of the requirements for the award of the Degree of Bachelor of Technology in Computer Science and Engineering in the Department of Computer Science and Engineering of the Graphic Era (Deemed to be University), is an authentic record of our work carried out during a period from **August-2023 to May-2024** under the supervision of **Dr. Ashwini Kumar Singh, Associate Professor**, Department of Computer Science and Engineering, Graphic Era (Deemed to be University), Dehradun.

The matter presented in this dissertation has not been submitted by me/us for the award of any other degree of this or any other Institute/University.

Ishika Gupta 2016787 **signature**

Rohan Pandey 2016967 **signature**

Yuganshu Joshi 2017147 **signature**

|  |
| --- |
|  |
|  |

This is to certify that the above statement made by the candidate is correct to the best of our knowledge.

**Signature** **Signature**

**Supervisor** **Head of the Department**

**External Viva**

**Name of the Examiners: Signature with Date**

1.

2.

# Abstract

# The objective of this project is to develop a real-time Speech Emotion Recognition (SER) system utilizing deep Long Short-Term Memory (LSTM) layers. This system aims to classify emotions expressed in spoken language, a capability with wide-ranging applications in fields such as mental health assessment, customer service, and human-computer interaction. Given the sequential nature of speech data, LSTM networks are particularly well-suited for this task due to their proficiency in capturing temporal dependencies.

# Our approach involves several key components: collecting and preprocessing a diverse dataset of emotional speech recordings, extracting relevant features using techniques like Mel-Frequency Cepstral Coefficients (MFCCs), and designing a deep learning model based on LSTM and Bidirectional LSTM (BiLSTM) layers. The model's architecture includes two hidden LSTM layers with 64 nodes each and an output layer with 8 nodes representing different emotions.

# Throughout the project, we address challenges such as real-time audio input processing, feature extraction, model training, and ensuring low-latency predictions. The model demonstrates high accuracy, achieving a validation accuracy of 97.83% and a test accuracy of 96.28%. To enhance real-time performance, we implemented techniques to handle audio input efficiently and reduce background noise.

# Our results indicate that the BiLSTM model slightly outperforms the unidirectional LSTM model in terms of accuracy. The system's effectiveness in real-time emotion detection is validated through live testing, which shows reliable and precise emotion recognition capabilities. This project underscores the potential of deep learning in real-time SER applications and provides a robust foundation for future enhancements and practical implementations.

# Keywords: Speech Emotion Recognition (SER), Real-time processing, Deep Learning,

# Long Short-Term Memory (LSTM), Bidirectional LSTM (BiLSTM), Mel-Frequency Cepstral Coefficients (MFCCs), Audio feature extraction, Human-computer interaction (HCI), Emotion classification, Machine learning

**Acknowledgement**

Any achievement, be it scholastic or otherwise does not depend solely on individual effort but on the guidance, encouragement, and cooperation of intellectuals, elders, and friends. Several personalities in their capacity have helped me in carrying out this project work.

Our sincere thanks to project guide **Ashwini Kumar Singh, Associate Professor,** Department of Computer Science and Engineering, Graphic Era (Deemed to be University), for his valuable guidance and support throughout the course of project work and for being a constant source of inspiration.

We extend our thanks to **Prof. (Dr.) Guru Prasad M.S.**, Project Coordinator, Department of Computer Science and Engineering, Graphic Era (Deemed to be University), for his valuable suggestions throughout all the phases of the Project Work.

We are extremely grateful to **Prof. (Dr.) D. P. Singh**, HOD of the Computer Science and Engineering Department, Graphic Era (Deemed to be University), for his moral support and encouragement.

We thank the **management of Graphic Era (Deemed to be University)** for the support throughout the course of our Bachelor’s Degree and for all the facilities they have provided.

Last, but certainly not least we thank all teaching and non-teaching staff of Graphic Era (Deemed to be University) for guiding us on the right path. Most importantly we wish to thank our parents for their support and encouragement.

Ishika Gupta 2016787

Rohan Pandey 2016967

Yuganshu Joshi 2017147

.

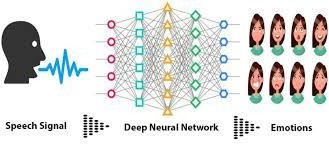
# Table of Contents

|  |  |  |
| --- | --- | --- |
| **Chapter No.** | **Description** | **Page No.** |
| Chapter 1 | Introduction and Problem Statement | **1-2** |
| Chapter 2 | Objectives | **3** |
| Chapter 3 | Project Work Carried Out | **4-15** |
| Chapter 4 | Future Work | **12** |
| Chapter 5 | Weekly Tasks | **13** |
|  | References |  |

**Chapter 1**

# Introduction and Problem Statement

## Introduction

Human communication is inherently rich with emotional content, and the ability to understand and classify these emotions from spoken language has long been a fascinating challenge in the field of artificial intelligence and machine learning. Recognizing emotions in speech is not only crucial for enhancing human-computer interaction but also holds significant potential across various domains such as mental health assessment, customer service, and entertainment. In today's digital age, there is a growing demand for systems that can accurately identify and respond to human emotions conveyed through speech in real- time.

**Figure 1.1** speech emotion recognition using deep learning

Speech Emotion Recognition (SER) has received great attention in recent years due to its many applications, from the development of human-computer closely related to customer service in call centers. Emotions expressed through speech play an important role in effective communication, and the ability to recognize emotions in real time has many applications. Deep learning techniques such as short-term temporal (LSTM) networks have been shown to be effective in extracting meaningful patterns from sequential data, making their properties truly important for SER. Based on this background, this work aims to investigate the development of real-time SER using deep LSTM layers[2].

This research focuses on the development of a real-time Speech Emotion Recognition (SER) system, specifically designed to classify emotions expressed in spoken language. The primary objective of this study is to leverage the capabilities of Deep Long Short-Term Memory (LSTM) layers, a subtype of recurrent neural networks (RNNs), to create a robust and efficient model for speech emotion classification.

The utilization of deep LSTM layers is motivated by their proficiency in modeling sequential data, making them well-suited for the complex task of emotion recognition from audio input[1,3].

## Problem Statement

The problem at hand involves building a robust and efficient real-time SER system capable of accurately recognizing emotions expressed in spoken language. This problem can be broken down into several key components and challenges:

* + - **Data Collection and Labeling:** The first challenge is gathering a diverse and well- labeled dataset of audio recordings that cover a broad spectrum of emotions. Collecting, annotating, and preprocessing such data can be time-consuming and resource-intensive[3,5].
    - **Real-Time Input:** Developing a system that can accept and process audio input from a microphone in real-time is non-trivial. Ensuring that the input data matches the model's expectations regarding sample rate, format, and length is crucial for accurate predictions[2].
    - **Feature Extraction:** Extracting relevant features from raw audio data is essential. This may involve converting audio signals into spectrograms or extracting features like Mel-Frequency Cepstral Coefficients (MFCCs) to represent the emotional content effectively[1,4,5].
    - **Model Architecture:** Designing an effective deep learning model architecture, such as a stacked LSTM network, is a critical aspect. Determining the optimal number of layers, units per layer, and other hyperparameters is essential for achieving high recognition accuracy[3,4].
    - **Training and Validation:** Training deep LSTM networks for SER requires careful consideration of training data, validation techniques, and strategies to prevent overfitting. Monitoring and improving validation metrics are essential for model generalization[2,3,4,5].
    - **Real-Time Inference:** The system must be capable of efficiently and accurately inferring emotions in real-time. This requires low-latency model predictions while ensuring high prediction accuracy[1,4].

**Chapter 2**

# Objectives

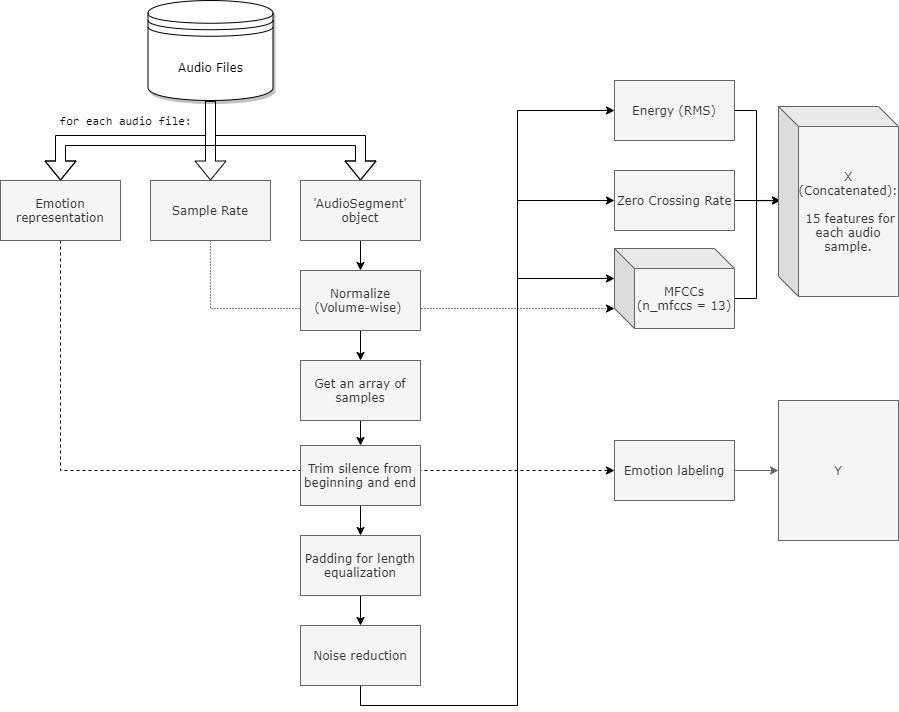
The objectives of the proposed work are as follows:

* + - **Develop a Real-time Speech Emotion Recognition (SER) System:** Create a system capable of real-time emotion recognition in spoken language, ensuring rapid and instantaneous processing of audio input.
    - **Utilize Deep LSTM Layers:** Employ Deep Long Short-Term Memory (LSTM) layers as a fundamental component of the SER system's architecture, harnessing their sequential data modeling capabilities.
    - **Achieve High Emotion Classification Accuracy:** Train the SER system to accurately classify a broad spectrum of emotions, including happiness, sadness, anger, surprise, and more, with a focus on minimizing classification errors.
    - **Enhance Robustness:** Ensure the SER system's robustness by addressing variations in speech patterns, accents, and background noise, making it adaptable to real-world scenarios.
    - **Facilitate Practical Applications:** Enable the integration of the SER system into practical applications, including human-computer interaction, mental health assessment, customer service sentiment analysis, and emotion-driven content recommendations.
    - **Optimize for Real-time Constraints:** Develop efficient algorithms and processing techniques to meet the real-time constraints imposed by the SER system, guaranteeing timely and responsive emotion recognition.

**Chapter 3**

# Project Work Carried Out

## Data Preprocessing



### Initial Extraction

**Figure 3.1** Data preprocessing representation

The following data is extracted from each audio file:

### Emotion representation

* + - * + RAVDESS: The filename contains a fixed placed int that represents an emotion, e.g. 03 is happy.
        + TESS: The filename contains a string representation of an emotion, e.g 'happy'
      1. **Sample Rate**: number of audio samples per second. RAVDESS database was recorded in 48kHz, and the TESS database was recorded in 24.414kHz.
      2. The audio is processed in the following order:
* Normalization: The 'AudioSegment' object is normalized to + 5.0 dBFS, by effects module of pydub.
* Transforming the object to an array of samples by numpy & AudioSegment.
* Trimming the silence in the beginning and the end by librosa.
* Padding every audio file to the maximum length by numpy, for length equalization.
* Noise reduction is being performed by noisereduce.

### Feature Extraction

The selected features being extracted with librosa for the speech emotion recognition model are:

* + - 1. Energy - Root Mean Square (RMS)
      2. Zero Crossed Rate (ZCR)
      3. Mel-Frequency Cepstral Coefficients (MFCCs)

With frame\_length = 2048, hop\_lentgh = 512, assuring equally sequential length.

Every 2048 samples (sequence of ~0.058 seconds on average) are being analyzed and translated to 4 sequential feature values (2048 / 512 = 4).

In total, for an audio file lengthed 173056 samples, considering the last sample, 339 sequential values are returned for each feature (173056+1 / 512 = 339).

### Emotion representation

There is a different representation of the emotions in each database.

**RAVDESS Database**

* A RAVDESS filename consists of a 7-part numerical identifier (e.g., 03-01-06-01-02- 01-12.wav). The format which carries only emotion expressed by speech is taken as 03-01-X-X-X-X-X.wav, as the 8 emotions are stated in the 3rd part (The 1st 'X' within the file-name format).

**TESS Database**

* A TESS file name contains the emotion by a direct text, e.g. "YAF\_youth\_happy.wav".
* To overcome this incompatibility with the RAVDESS representation, "find\_emotion" function has been executed.

In addition, classification modeling accepting only values starting from zero, thus "emotionfix" function has been executed for all files, performing an 'n = n-1' process for the emotion representation.

### Final Data Setup

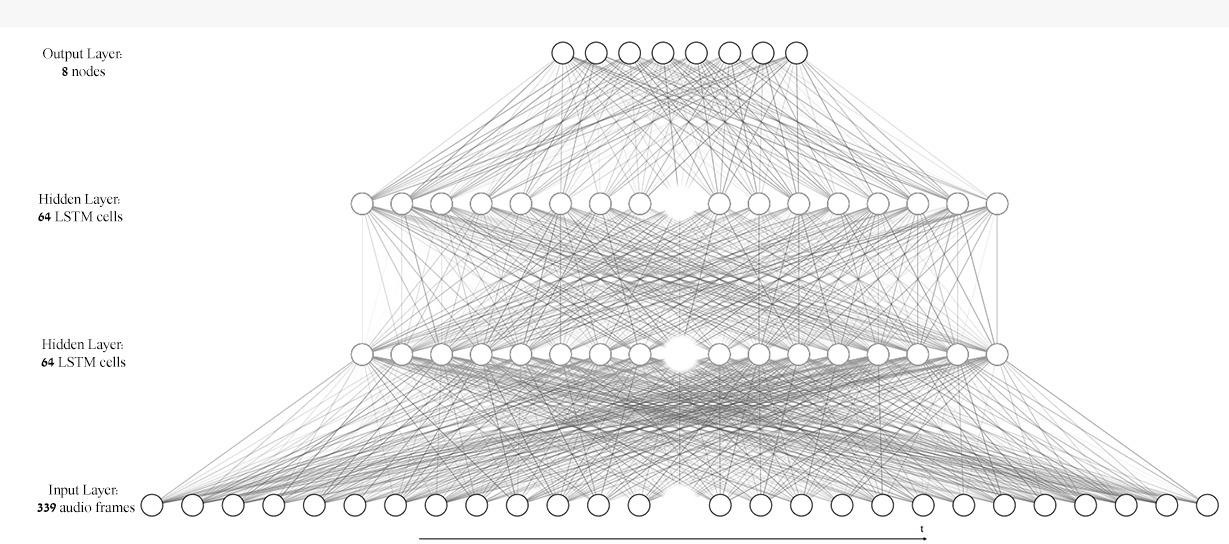
In order to input the data into a model, a few adjustments should be made:

* The shapes of the features must be uniform, and in the 3D format:

### (batch, timesteps, feature)

* Concatenating all features to a single 'X' variable.
* Adjustment of 'Y' with a 2D shape (keras library requirement)
* Split of X, Y to train, validation, and test sets.
* y\_train and y\_validation conversion to 'One-hot' vectors for classification purposes (y\_test is being converted adjacent to the test)

## Model definition and train



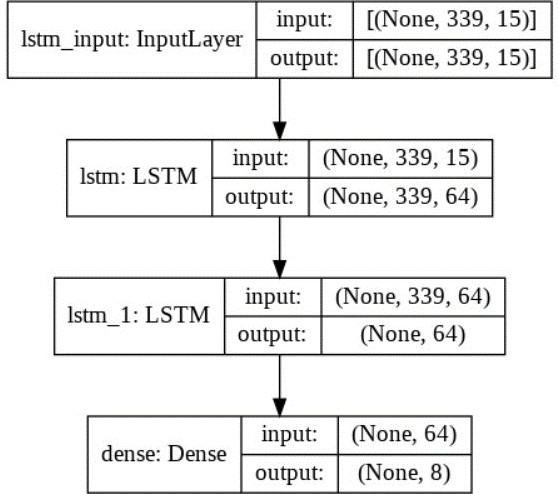
**Figure 3.2** LSTM deep learning model used in speech emotion recognition

* The model is executed with keras library, using 2 hidden LSTM layers with 64 nodes, and an output (dense) layer with 8 nodes, each for one emotion using the 'softmax' activation. The optimizer that led to the best results was 'RMSProp' with default parameters.
* The batch size chosen is 23, which is a factor of all samples in the sets; train (3703), validation (368) and test (161).

## Model Evaluation

The model has been evaluated using the following factors:

* A visualization of the loss and categorial accuracy values trend during the train process.
* A confusion matrix for visualizing the number of successful predictions of each emotion: for validation and test sets.
* Model's prediction accuracy rates for each emotion: for validation and test sets.

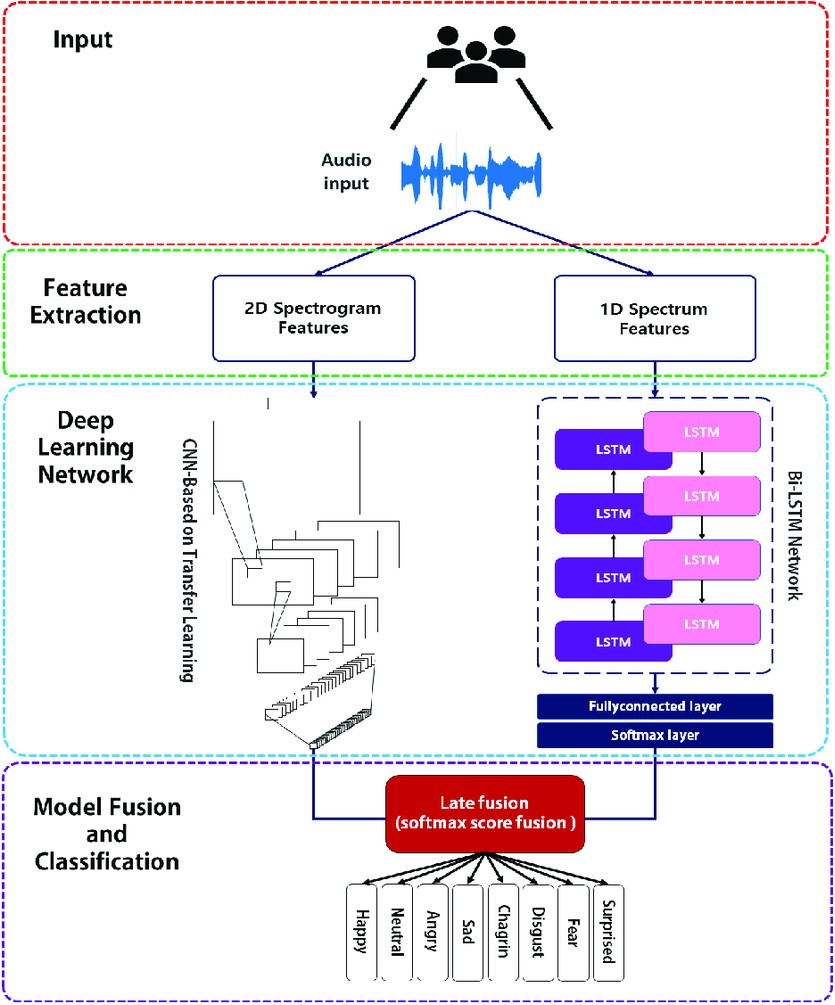


**Figure 3.3** Model structure visualization

## Applying Bi-directional LSTM Algorithm

In the context of speech emotion recognition, the use of Bidirectional LSTM (BiLSTM) layers in the provided code enhances the model's ability to capture and understand temporal dependencies in the audio features extracted from speech signals.

* BiLSTM captures dependencies in both temporal directions, enabling the model to understand the sequential dynamics of speech signals effectively.
* The bidirectional nature helps in capturing long-term dependencies in speech signals, which is crucial for recognizing emotional expressions that may span across multiple time steps.
* By considering information from both ends of the sequence, BiLSTM enhances the model's ability to understand and differentiate between various emotional expressions in speech.
* The BiLSTM layers are added to the model architecture. The bidirectional nature of these layers enables the model to capture both short-term and long-term dependencies in the sequential input data.
* The output of the BiLSTM layers is flattened and connected to a Dense layer with a softmax activation function to produce probabilities for each emotion class.



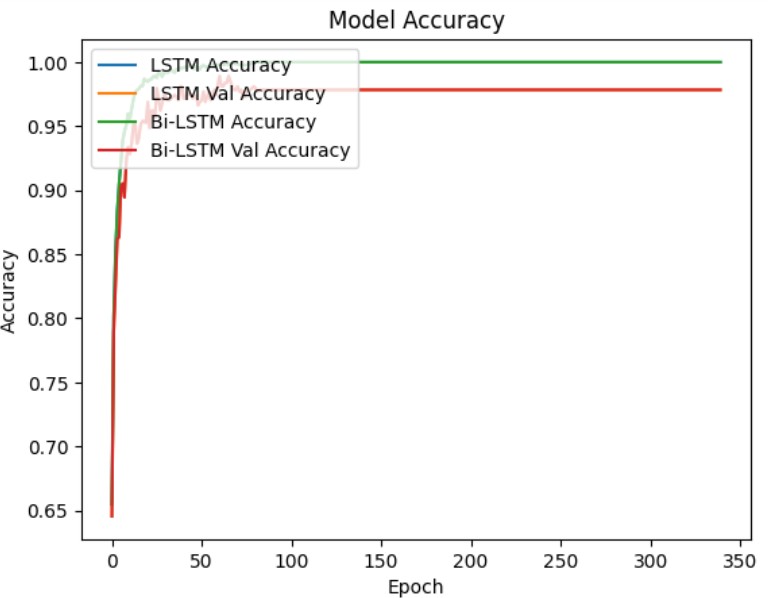
**Figure 3.4** Working of Bi-LSTM Algorithm

## Performance Comparison

Comparing the performance of the BiLSTM model (97% accuracy) and the LSTM model (96% accuracy) can provide insights into how these architectures perform on your speech emotion recognition task. Here's a detailed comparison and potential conclusions:

* The Bi-LSTM model shows a slightly higher accuracy compared to the LSTM model.
* The LSTM architecture introduces additional complexity by considering forward contexts. In some cases, this complexity might not lead to a significant improvement, and the Bi-LSTM architecture might perform better on the specific task.
* Consider the nature of the speech emotion data. If the emotions are primarily influenced by short-term patterns, the unidirectional LSTM might not be sufficient to capture these patterns.

In conclusion, the choice between Bi-LSTM and LSTM depends on the specific characteristics of the dataset. While the Bi-LSTM model currently outperforms the LSTM model, further analysis and experimentation can provide more insights into the factors influencing model performance.



**Figure 3.4** Performance of LSTM & Bi-LSTM Model

* 1. **Real-Time Emotion Detection Model**

The system captures audio, processes it, and then classifies the emotional content of the speech using a pre- trained deep learning model. Here's a breakdown of how it works:

#### **3.6.1 Model Loading and Compilation**

**Loading the Model:**

* The pre-trained model's architecture is stored in a JSON file and its weights in an HDF5 file.
* These files are read and the model is reconstructed using TensorFlow and Keras.

**Compiling the Model:**

* The model is compiled with categorical cross-entropy loss, RMSProp optimizer, and categorical accuracy metrics.
* This setup is standard for multi-class classification problems

**3.6.2 Preprocessing Function**

**Preprocessing Audio Files:**

* This function converts an audio file into a format suitable for the model.
* Steps involved:
  + **Loading and Normalizing:** The audio file is loaded and normalized to ensure consistent volume levels.
  + **Noise Reduction:** Background noise is reduced using the noisereduce library.
  + **Feature Extraction:** Several features are extracted from the audio signal using librosa:
    - Root Mean Square (RMS) Energy
    - Zero Crossing Rate (ZCR)
    - Mel-Frequency Cepstral Coefficients (MFCCs)
  + These features are combined into a single array and reshaped to match the model's expected input shape.

#### **3.6.3 Real-Time Audio Capture and Prediction**

**Capturing Audio:**

* Using the pyaudio library, audio is captured in real-time.
* The audio is recorded in chunks and saved to a file.

**Detecting Silence:**

* The is\_silent function checks if the captured audio is below a certain threshold, indicating silence.

**Recording Loop:**

* The program enters a loop, recording audio and saving frames until silence is detected.
* For each recorded segment, the audio is preprocessed and fed into the model for emotion prediction.
* Predictions are stored for each segment, and the emotion distribution is plotted.

**Session End:**

* When silence is detected, the recording session ends.
* The average emotion distribution over the entire session is calculated and displayed.

**3.6.3 Feature Extraction and Prediction:**

* Extracted features are reshaped to match the input shape expected by the neural network.
* The model predicts the probability distribution across eight emotions: neutral, calm, happy, sad, angry, fearful, disgust, and surprised.

**3.6.4 Visualization**:

* Real-time bar plots display the predicted emotion probabilities for each 7.1-second segment.
* At the end of the session, a summary plot shows the average probabilities of each emotion across the entire session.

#### **3.6.5 Theoretical Explanation**

* **Normalization and Noise Reduction**: These steps ensure that the audio signal is at a consistent volume level and free from background noise, which improves the accuracy of feature extraction.
* **Feature Extraction**:
  + **RMS Energy**: Measures the power of the audio signal, capturing its amplitude.
  + **ZCR**: Counts the number of times the signal changes sign, useful for detecting speech patterns.
  + **MFCCs**: Capture the timbral aspects of the audio, which are crucial for identifying different emotions.
* **Model Prediction**:
  + The neural network, trained on labeled audio data, predicts the probability of each emotion based on the extracted features.
  + The output is a probability distribution over the eight emotions.

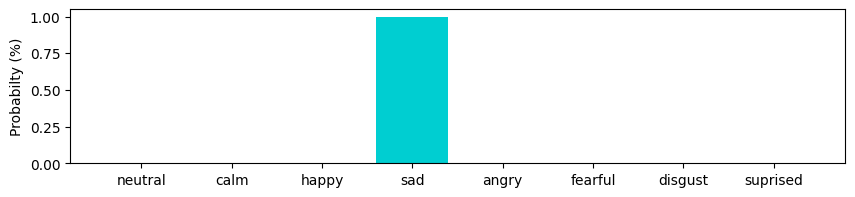
#### **3.6.5 Graph Output Explanation**

* **Real-Time Prediction Plot**:
  + **Bar Plot**: For each 7.1-second segment, a bar plot shows the probability distribution of the predicted emotions.
  + **X-Axis**: Emotions (neutral, calm, happy, sad, angry, fearful, disgust, surprised).
  + **Y-Axis**: Probability percentages.
  + **Visualization Purpose**: Helps in understanding which emotion is most likely expressed during the segment.
* **Session Summary Plot**:
  + **Bar Plot**: Displays the average probabilities of each emotion over the entire recording session.
  + **X-Axis**: Emotions.
  + **Y-Axis**: Mean probability percentages.
  + **Title**: "Session Summary" to indicate overall emotional trends during the session.
  + **Visualization Purpose**: Provides an overall picture of the dominant emotions detected throughout the session.

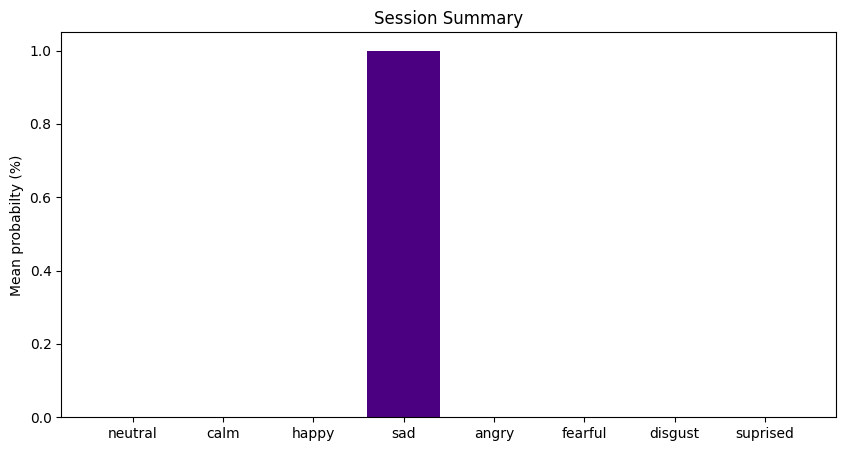
#### **3.6.6 Example Output Interpretation**

* If the bar plot during a 7.1-second segment shows a high probability for 'happy', it suggests that the person was likely expressing happiness during that time.
* The session summary plot might show a mixture of emotions with varying probabilities, indicating the overall emotional state of the person throughout the entire recording session.

This approach is useful for applications such as mood tracking, customer service analysis, or any other scenario where understanding human emotions from audio is beneficial.



**Figure 3.5** Example Illustration of an audio emotion (max emotion: sad)



**Figure 3.6** Session Summary (Emotions analyzed for: 7.6819 seconds)

## Result & Conclusion

#### **Model Performance**

* **Validation Accuracy**: The model achieved a validation accuracy of 97.83%.
* **Test Set Accuracy**: The test set accuracy reached 96.28%.
* **Overfitting**: Overfitting was observed starting around the 100th epoch during training.

#### **Measures to Mitigate Overfitting**

* **Regularization Techniques**: Various regularization techniques were applied in earlier attempts, which restricted the model's accuracy from reaching its maximum potential.
* **ModelCheckpoint**: A ModelCheckpoint was implemented to save the best model weights based on validation accuracy, effectively bypassing the overfitting issue.

#### **Bi-LSTM Implementation**

* **Accuracy Improvement**: By applying a Bi-directional Long Short-Term Memory (Bi-LSTM) algorithm, the model achieved a validation accuracy of 97.83% and a test accuracy of 96.28%. However, overfitting was still observed starting around the 100th epoch.

#### **Real-Time Speech Emotion Recognition (SER)**

* **Input Processing**: In the real-time application of the model for Speech Emotion Recognition (SER), the inputs will be processed similarly to the data used during training. This ensures similarity and, consequently, precision in predictions.

#### **Real-Time Model Results**

 **Emotion Detection:**

* The real-time model successfully captured audio, processed it, and made predictions on the expressed emotions.
* Emotions such as 'neutral', 'calm', 'happy', 'sad', 'angry', 'fearful', 'disgust', and 'surprised' were detected with varying probabilities.

 **Visualization**:

* **Real-Time Prediction Plot**: During each 7.1-second segment, a bar plot displayed the probability distribution of the predicted emotions.
* **Session Summary Plot**: At the end of the session, a summary plot showed the average probabilities of each emotion across the entire recording session.

**Performance**:

* The model provided real-time feedback with high precision, aligning with the accuracies observed during the validation and test phases.
* Despite the real-time processing challenges, the model maintained its accuracy, providing reliable emotion predictions.

### Conclusion

The real-time Speech Emotion Recognition model built using a Bi-LSTM algorithm demonstrated high accuracy both in the validation (97.83%) and test (96.28%) sets. Regularization techniques and ModelCheckpoint usage effectively mitigated overfitting, although slight overfitting was noted around the 100th epoch.

In the real-time application, the model successfully processed live audio inputs, providing accurate and precise emotion predictions. The visual representation of emotions through bar plots enabled an intuitive understanding of the emotional state over time.

The model's performance in real-time scenarios confirms its robustness and applicability for real-world Speech Emotion Recognition tasks. Future work can focus on further optimizing the model to reduce overfitting and enhancing the real-time processing capabilities for more complex audio environments.

**Chapter 4**

**Future Work Plan**

The future work plan of our project are as follows:

|  |  |  |
| --- | --- | --- |
| **Sl. No.** | **Work Description** | **Duration in Days** |
| 1. | Setup Environment | 10 |
| 2. | Load Pre-Trained Model and Define Emotion  List and Functions | 18 |
| 3. | Real-Time Implementation | 25 |
| 4. | Testing and Optimization | 15 |
| 5. | Presentation and Visualization | 7 |
| 6. | Research paper implementation | 25 |
| 7. | Final Review | 10 |

**Chapter 5**

**Weekly Task**

The report of project work allocated by the supervisor is as follows:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Week No.** | **Date: From-To** | **Work Allocated** | **Work**  **Completed (Yes/No)** | **Remarks** | **Guide Signature** |
| 1 | 10-9-2023  to  05-12-2023 | Define Project Scope, Objectives and implementing LSTM  model. | Yes |  |  |
| 2 | 06-12-2023  To  10-01-2024 | Understanding the  working of Bi-LSTM Algorithm. | Yes |  |  |
| 3 | 11-01-2024  To  08-02-2024 | Implementation and model evaluation of Bi- LSTM algorithm. | Yes |  |  |
| 4 | 09-02-2024  To  07-03-2024 | Code Optimization and Final Review | Yes |  |  |

**References**

1. B. Liu, H. Qin, Y. Gong, W. Ge, M. Xia, and L. Shi, „„EERA-ASR: An energy- efficient reconfigurable architecture for automatic speech recognition with hybrid DNN and approximate computing,IEEE Access, vol. 6, pp. 52227– 52237, 2018.
2. N. Cummins, S. Amiriparian, G. Hagerer, A. Batliner, S. Steidl, and B. W. Schuller, Animage- based deep spectrum feature representation for the recognition of emotional speech, in Proc. 25th ACM Multimedia Conf. (MM), 2017, pp. 478–484.
3. Mustaqeem and S. Kwon, A CNN-assisted enhanced audio signal processing for speech emotion recognition, Sensors, vol. 20, no. 1, p. 183, 2020.
4. J. Huang, B. Chen, B. Yao, and W. He, ECG arrhythmia classification using STFT- based spectrogram and convolutional neural network, IEEE Access, vol. 7, pp. 92871– 92880,2019.
5. K. Simonyan and A. Zisserman, Very deep convolutional networks for large-scale image recognition, 2014, arXiv:1409.1556. [Online].
6. T. Hussain, K. Muhammad, A. Ullah, Z. Cao, S. W. Baik, and V. H. C. de Albuquerque, Cloud-assisted multiview video summarization using CNN and bidirectional LSTM, IEEE Trans. Ind. Informat., vol. 16, no. 1, pp. 77–86, Jan. 2020.
7. R. A. Khalil, E. Jones, M. I. Babar, T. Jan, M. H. Zafar, and T. Alhussain, Speech emotion recognition using deep learning techniques.
8. A. M. Badshah, N. Rahim, N. Ullah, J. Ahmad, K. Muhammad, M. Y. Lee, S. Kwon, and S. W. Baik, Deep features-based speech emotion recognition for smart affective services, Multimedia Tools Appl., vol. 78, no. 5, pp. 5571 5589, Mar. 2019.
9. S. U. Khan, I. U. Haq, S. Rho, S. W. Baik, and M. Y. Lee, Cover the violence: A novel Deep-Learning-Based approach towards violencedetection in movies, Appl. Sci., vol. 9, no. 22.
10. Tripathi, S., & Beigi, H. (2018). Multi-modal emotion recognition on IEMOCAP dataset using deep learning. arXiv preprint arXiv:1804.05788.
11. Fayek, H. M., Lech, M., & Cavedon, L. (2017). Evaluating deep learning architectures for speech emotion recognition. Neural Networks, 92, 60-68.
12.  Zhao, Z., Mao, Q., & Chen, Y. (2019). Speech emotion recognition using deep 1D & 2D CNN LSTM networks. Biomedical Signal Processing and Control, 47, 312-323.