**A SYNOPSIS 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**

**September-2023**

GEU logo

**CANDIDATE’S DECLARATION**

I/We hereby certify that the work which is being presented in the Synopsis 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), Dehradun shall be carried out by the undersigned under the supervision of **Dr. Ashwini Kumar Singh, Associate Professor**, Department of Computer Science and Engineering, Graphic Era (Deemed to be University), Dehradun.

Ishika Gupta 2016787 signature

Rohan Pandey 2016967 signature

Yuganshu Joshi 2017147 signature

The above mentioned students shall be working under the supervision of the undersigned on the **“Speech Emotion Recognition (Classification) in real-time using Deep LSTM layers”**

Signature Signature

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

**Internal Evaluation (By DPRC Committee)**

**Status of the Synopsis:** Accepted / Rejected

**Any Comments:**

**Name of the Committee Members: Signature with Date**

1.

2.

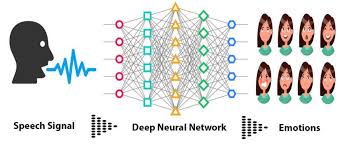
**Table of Contents**

| **Chapter No.** | **Description** | **Page No.** |
| --- | --- | --- |
| Chapter 1 | Introduction and Problem Statement | **1-2** |
| Chapter 2 | Background/ Literature Survey | **3-4** |
| Chapter 3 | Objectives | **5** |
| Chapter 4 | Hardware and Software Requirements | **6** |
| Chapter 5 | Possible Approach/ Algorithms | **7-9** |
|  | References |  |

**Chapter 1**

**Introduction and Problem Statement**

* 1. **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].

* 1. **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**

**Background/ Literature Survey**

A literature survey or background research on speech emotion recognition (SER) using deep LSTM layers provides valuable insights into existing methods, challenges, and developments in this field. Here is an overview of some key research findings and trends in the literature :

* **Deep Learning and SER:**

Deep learning techniques, including recurrent neural networks (RNNs) like LSTM, have demonstrated remarkable success in SER. LSTMs are favored for their ability to model temporal dependencies in sequential data, making them suitable for processing audio signals[3].

* **Datasets:**

Several publicly available datasets have played a crucial role in advancing SER research. Examples include the IEMOCAP, EmoReact, and RAVDESS datasets, each containing a variety of emotional expressions[3].

* **Feature Extraction:**

Traditional feature extraction techniques, such as Mel-Frequency Cepstral Coefficients (MFCCs), chroma features, and spectral features, have been used alongside deep learning models to capture relevant acoustic characteristics[4,5].

* **Model Architectures:**

Deep neural network architectures for SER often involve combinations of LSTM layers with other components like convolutional neural networks (CNNs) or attention mechanisms.Attention mechanisms have been employed to focus on specific segments of audio data for improved emotion recognition[2].

* **Transfer Learning:**

Transfer learning, using pre-trained models on large datasets (e.g., ImageNet), has been explored to improve SER performance when limited emotion-specific data is available[1].

* **Real-Time SER:**

Developing real-time SER systems has been an emerging trend to enable applications like emotion-aware virtual assistants and emotion recognition in video conferencing. Latency reduction and efficient model architectures are crucial considerations for real-time applications[1,4].

* **Challenges:**

Challenges in SER research include dealing with noisy data, variations in emotional expression, and the need for robustness across different languages and cultures.Model overfitting and the scarcity of labeled emotional speech data for certain emotions are persistent challenges[2].

* **Evaluation Metrics:**

Common evaluation metrics in SER research include accuracy, F1-score, and confusion matrices. However, the choice of evaluation metrics may vary depending on the specific application[2,4,5].

* **Applications:**

SER finds applications in diverse domains, including human-computer interaction, call center analytics, mental health monitoring, and affective computing.Its potential for enhancing user experiences in virtual reality (VR) and gaming has also been explored[1,3].

* **Ethical Considerations:**

Privacy and ethical concerns related to the collection and use of emotional data have gained attention in SER research. Ensuring user consent and data security are important aspects[5].

* **Future Directions:**

Future research in SER is expected to focus on improving the robustness of models across different languages and cultures, reducing the need for large labeled datasets through semi-supervised or unsupervised learning, and developing more interpretable models[5] .

**Chapter 3**

**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 4**

**Hardware and Software Requirements**

4.1 Hardware Requirements

| Sl. No | Name of the Hardware | Specification |
| --- | --- | --- |
| 1. | Computer or Server | A quality microphone connected to your computer is necessary. |
| 2. | GPU | GPU memory of 4GB is sufficient for entry-level deep learning models. 8GB memory is recommended for optimal results. |
| 3. | Memory (RAM) | Minimum of 16GB RAM |
| 4. | Storage | Adequate storage space is required for storing your dataset, model checkpoints, and other project-related files |
| 5. | Microphone | A quality microphone connected to your computer is necessary. |

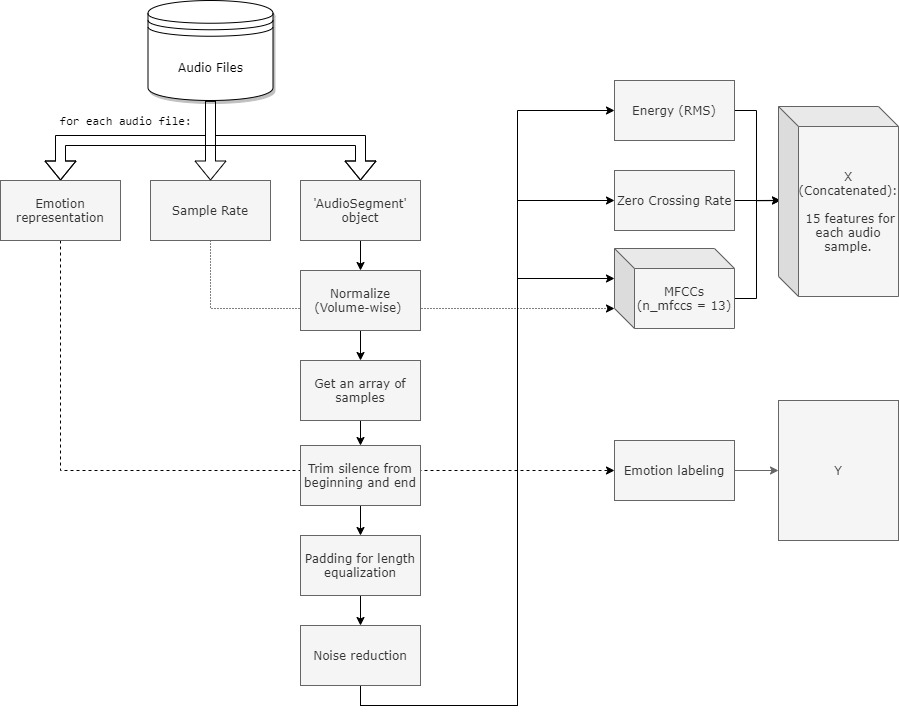
4.2 Software Requirements

| Sl. No | Name of the Software | Specification |
| --- | --- | --- |
| 1. | Operating System | Preferred Operating Systems are   * Windows (version 10 or higher) * Linux (Ubuntu) |
| 2. | Python | Need Python (version 3.6 or later) with necessary packages as Python is commonly used for deep learning and computer vision projects. |
| 3. | Integrated Environment | IDE is required to provide a unified interface to write, edit, test and debug code (I.e. Visual studio Code, PyCharm) |
| 4. | Libraries and Packages | TensorFlow(for backend functioning)  Pydub for audio segment and effects.  Libarosa,noisereduce and keras . |

**Chapter 5**

**Possible Approach/ Algorithms**

**5.1 Data Preprocessing**

****

**Figure 5.1** Data preprocessing representation

In general, The following data is extracted from each audio file:

1. **Emotion representation**
   * RAVDESS: The filename contains a number that represents an emotion, e.g. 03 is happy.
   * TESS: The filename contains a direct representation of an emotion, e.g 'happy'.
2. **Sample Rate**: number of audio samples per second. RAVDESS database was recorded in 48kHz, and the TESS database was recorded in 22.5kHz.
3. **The audio** is processed in the following order:

* 'AudioSegment' instance: The audio is loaded to an object by the library 'AudioSegment'.
* Normalization: Each 'AudioSegment' object is normalized to + 5.0 dBFS.
* Transforming the object to an array of samples is crucial for the rest of the preprocessing.
* Trimming the silence in the beginning and the end, to get rid of unnecessary data.
* Padding every audio file to the same length, for length equalization.
* Noise reduction is being performed.

**5.2 Features extraction**

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

**Energy - Root Mean Square (RMS)**

**Zero Crossed Rate (ZCR)**

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).

**5.3 Emotion representation**

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

**RAVDESS Database**

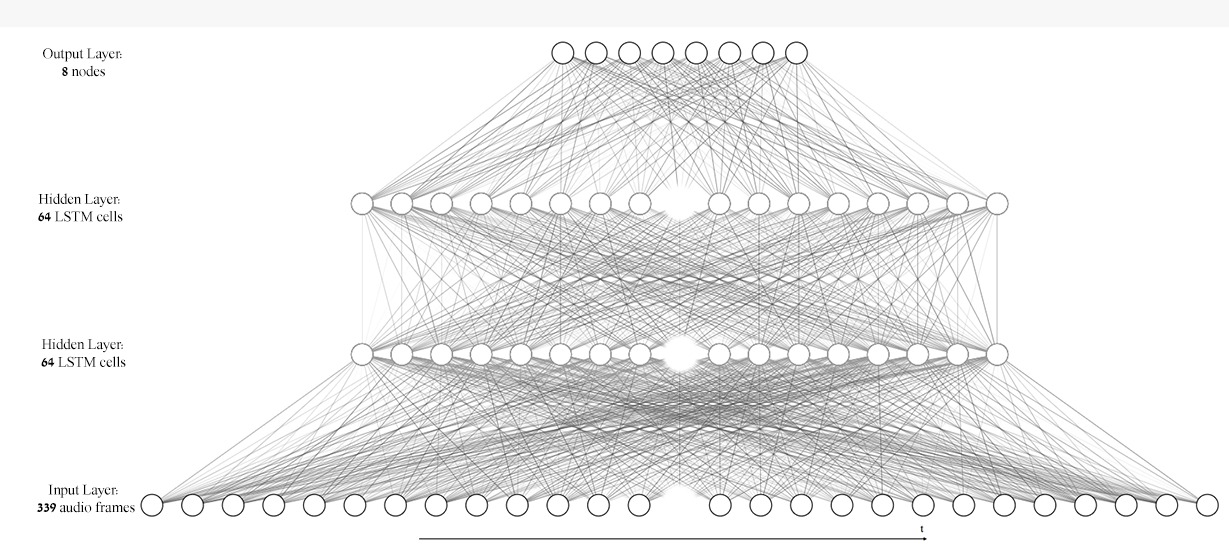
A RAVDESS filename consists of a 7-part numerical identifier (e.g., 03-01-06-01-02-01-12.wav). The format which carries only emotion expressed by speech is taken as 03-01-X-X-X-X-X.wav, as the 8 emotions are stated in the 3rd part (The 1st 'X' within the file-name format).

**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, performing an 'n = n-1' process for the emotion representation

**5.3 Model definition and train**

****

**Figure 5.2** 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).

**References**

[1] R.LEELAVATHI1, S.ARUNA DEEPTHI2, V.ARUNA3, “SPEECH EMOTION RECOGNITION USING LSTM” Volume: 09 Issue: 01 | Jan 2022, p-ISSN: 2395-0072[2] Deep Learning Techniques for Speech Emotion Recognition, from Databases to Models [Online]. Accessed on 5th September 2023: <https://www.mdpi.com/1424-8220/21/4/1249>

[3] An ongoing review of speech emotion recognition [Online]. Accessed on 8th September 2023: <https://www.sciencedirect.com/science/article/pii/S0925231223000103>

[4] ABDELAZIZ A. ABDELHAMID 1,2, EL-SAYED M. EL-KENAWY 3,4, (Senior Member, IEEE), BANDAR ALOTAIBI 5,6, (Member, IEEE), GHADA M. AMER7 , MAHMOUD Y. ABDELKADER1 , ABDELHAMEED IBRAHIM 8 , (Member IEEE), AND MARWA METWALLY EID 4, “Robust Speech Emotion Recognition Using CNN+LSTM Based on Stochastic Fractal Search Optimization Algorithm” Received March 16, 2022, accepted April 27, 2022, date of publication May 5, 2022, date of current version May 12, 2022.

[5] Speech Emotion Recognition (SER)Using CNN And LSTMs [Online]. Accessed on 8th September 2023: <https://pub.towardsai.net/speech-emotion-recognition-ser-using-cnn-and-lstms-4a5dc4c314fd>