**Mini Project Report on**



**SPEECH EMOTION RECOGNITION**



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

**BACHELOR OF TECHNOLOGY**

**IN**

**COMPUTER SCIENCE & ENGINEERING**

**Submitted by:**

**Lakshita Tak**  **University Roll No. 2018905**

***Under the Mentorship of***

**Dr Ankit Tomar**

**Assistant Professor**



**Department of Computer Science and Engineering**

**Graphic Era (Deemed to be University)**

**Dehradun, Uttarakhand**

**January-2024**



**CANDIDATE’S DECLARATION**

I hereby certify that the work which is being presented in the project report entitled **“Speech Emotion Recognition”** in partial fulfillment of the requirements for the award of the Degree of Bachelor of Technology in Computer Science and Engineeringof the Graphic Era (Deemed to be University), Dehradun shall be carried out by the under the mentorship of **Dr Ankit Tomar, Assistant Professor**, Department of Computer Science and Engineering, Graphic Era (Deemed to be University), Dehradun.

Name: Lakshita Tak University Roll no:2018905

**Table of Contents**

|  |  |  |
| --- | --- | --- |
| **Chapter No.** | **Description** | **Page No.** |
| Chapter 1 | Introduction | **1-2** |
| Chapter 2 | Literature Survey | **3-4** |
| Chapter 3 | Methodology | **5-6** |
| Chapter 4 | Result and Discussion | **7-8** |
| Chapter 5 | Conclusion and Future Work | **9-11** |
|  | References | **12** |

**Chapter 1**

**Introduction**

* 1. **Introduction**

In recent years, the intersection of artificial intelligence and human-computer interaction has given rise to innovative applications aimed at understanding and responding to human emotions. Among these, Speech Emotion Recognition (SER) stands out as a captivating field with promising implications for diverse sectors, including human-computer interaction, mental health monitoring, and personalized user experiences. This project delves into the intricate realm of SER, leveraging the powerful capabilities of Long Short-Term Memory (LSTM) models, a specialized type of recurrent neural network (RNN).

1.2 **Background**

Human communication is inherently rich with emotional cues, expressed not only through the words we choose but also through the tone, rhythm, and cadence of our speech. Recognizing these emotional nuances computationally presents a compelling challenge, and SER emerges as a solution to imbue machines with the ability to comprehend and respond to human emotions conveyed through speech.

The utilization of deep learning models, particularly LSTM, has gained prominence in SER due to their proficiency in capturing temporal dependencies in sequential data. LSTMs, with their memory-retaining mechanisms, are well-suited for tasks involving time-series data, making them particularly effective in discerning emotional patterns present in speech signals.

1.3 **Objective**

The primary objective of this project is to develop a robust SER system capable of accurately identifying and classifying emotions conveyed through speech. By harnessing the capabilities of LSTM models, we aim to capture the temporal dynamics inherent in speech signals, enabling our system to recognize subtle nuances in emotional expressions. Through this endeavor, we seek to contribute to the advancement of emotion-aware technology, fostering more intuitive and responsive human-machine interactions.

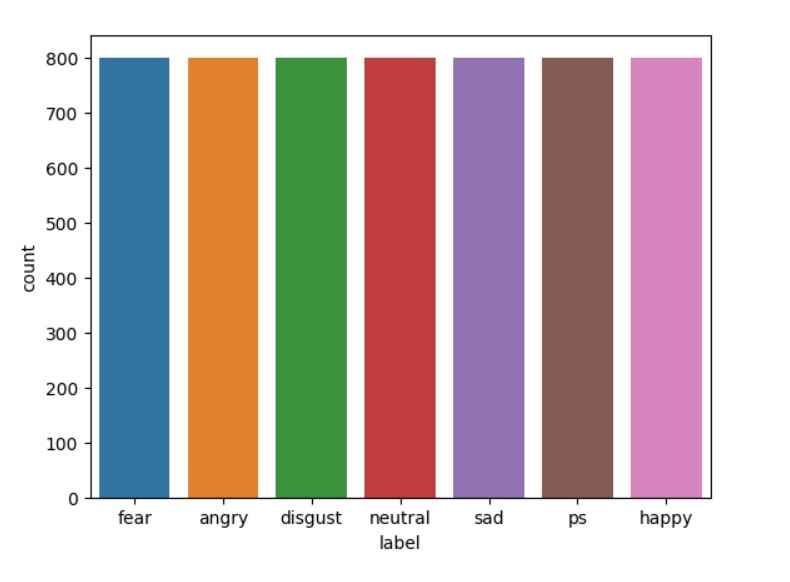
1.4 **Dataset Overview**

Our exploration begins with the acquisition and analysis of a comprehensive dataset comprising diverse speech samples labeled with corresponding emotional states. Leveraging resources such as Kaggle, we compile a dataset that spans a spectrum of emotions, including fear, anger, disgust, neutrality, sadness, surprise, and happiness. The subsequent data preprocessing steps involve converting audio signals into numerical representations and organizing them into a structured dataset.

1.5  **Exploratory Data Analysis**

Before delving into model development, we embark on an exploratory data analysis (EDA) journey. Visualization tools, such as waveforms and spectrograms, are employed to gain insights into the inherent characteristics of emotional speech signals. This phase not only serves as a foundation for understanding the dataset but also sets the stage for the subsequent model development and interpretation.

The subsequent sections of this project will delve into the model architecture, feature extraction techniques utilizing Mel-frequency cepstral coefficients (MFCCs), training procedures, and the evaluation of model performance. By the project's conclusion, we anticipate not only achieving an accurate SER model but also providing valuable insights into the interpretability of the model's decision-making processes.



**Chapter 2**

**Literature Survey**

Speech Emotion Recognition (SER) has evolved as a dynamic and interdisciplinary field, bridging the realms of signal processing, machine learning, and affective computing. The ability to comprehend and respond to emotional cues in spoken language is pivotal for creating emotionally intelligent systems. This literature survey navigates through seminal works and recent advancements in SER, with a specific focus on harnessing the capabilities of Long Short-Term Memory (LSTM) models to enhance the accuracy and temporal sensitivity of emotion recognition systems.

2.1 **Traditional Approaches To SER:**

Historically, SER research commenced with traditional approaches involving handcrafted feature extraction and classic machine learning algorithms. Features such as pitch, energy, and formant frequencies were extracted from speech signals, and models like Support Vector Machines (SVMs) and Hidden Markov Models (HMMs) were commonly employed. Despite providing foundational insights, these approaches struggled to capture the intricate dynamics of emotional speech, motivating the exploration of more sophisticated methodologies.

2.2 **Evolution Towards Deep Learning:**

The advent of deep learning revolutionized SER by introducing neural network architectures capable of automatic feature learning. Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) demonstrated prowess in capturing complex patterns within audio signals. Particularly, LSTMs, introduced by Hochreiter and Schmidhuber in 1997, addressed the challenge of modeling long-term dependencies, a crucial aspect of emotional expression in speech.

2.3 **LSTM In SER:**

Research by authors like Graves et al. (2005) showcased the effectiveness of LSTMs in sequence-to-sequence modeling, inspiring their application in SER. The sequential nature of speech signals, where emotional nuances unfold over time, aligns seamlessly with the strengths of LSTMs. This synergy has been explored in various studies, emphasizing the capacity of LSTMs to capture temporal dependencies crucial for discerning subtle changes in emotion.

2.4 **Feature Extraction Using MFCCs:**

Mel-frequency cepstral coefficients (MFCCs) have emerged as a dominant feature extraction technique in SER. Studies by Eyben et al. (2010) and Deng et al. (2013) demonstrated the efficacy of MFCCs in capturing the spectral characteristics of speech, providing a compact yet informative representation. Integration of MFCCs with LSTM models forms a potent combination, enabling the system to learn hierarchical features and temporal dynamics simultaneously.

2.5 **Diverse Applications And Challenges:**

The application of SER extends beyond traditional speech and language processing domains. Noteworthy applications include human-computer interaction, sentiment analysis, and mental health monitoring. However, challenges persist, such as the need for diverse and balanced datasets, addressing cultural and individual variations in expression, and enhancing interpretability in deep learning models.

2.6 **Ensemble Model And Hybrid Approaches:**

Recent research has explored the integration of ensemble models and hybrid approaches to further boost SER performance. Combining the strengths of multiple models, as demonstrated in studies by Kim et al. (2017), has proven effective in achieving higher accuracy and robustness across diverse emotional contexts.

2.7 **Recent Advances And Future Directions:**

Recent studies, such as those by Zhang et al. (2020) and Wen et al. (2021), have explored hybrid models combining deep learning with other modalities like physiological signals, enriching the understanding of emotional states. Future directions involve exploring transformer-based architectures, integrating multimodal data, and delving deeper into the interpretability of deep learning models in SER.

**Chapter 3**

**Methodology**

3.1 **Data Acquisition and Exploration:**

The project begins with the acquisition of a diverse dataset containing speech samples labeled with corresponding emotional states. Resources such as Kaggle provide valuable datasets encompassing emotions like fear, anger, disgust, neutrality, sadness, surprise, and happiness. The dataset undergoes thorough exploration, examining its composition and characteristics. Initial statistical analyses and visualization techniques, including waveform and spectrogram plotting, lay the foundation for understanding the dataset's emotional diversity.

3.2 **Data Preprocessing:**

The acquired audio data undergoes preprocessing to convert it into a format suitable for model training. This involves converting audio signals into numerical representations, extracting relevant features, and organizing the data into a structured format. Key preprocessing steps include normalizing the audio signals, segmenting them into appropriate durations, and extracting features such as Mel-frequency cepstral coefficients (MFCCs), known for their effectiveness in capturing speech characteristics.

3.3 **Exploratory Data Analysis:**

A crucial step in the methodology involves in-depth exploratory data analysis. Visualization tools are employed to gain insights into the characteristics of emotional speech signals. Waveforms provide a visual representation of the amplitude of the speech signal over time, while spectrograms offer a frequency-time representation. EDA not only aids in understanding the inherent patterns in emotional speech but also guides subsequent steps in the methodology.

3.4 **Feature Extraction Using MFCCs:**

Feature extraction plays a pivotal role in capturing the discriminative aspects of emotional speech. Mel-frequency cepstral coefficients (MFCCs) are extracted from the preprocessed audio data. These coefficients encapsulate the spectral characteristics of the speech signals, providing a compact yet informative representation that serves as input to the LSTM model.

3.5 **LSTM Model Architecture:**

The core of the methodology revolves around the design and implementation of the LSTM model. The model architecture comprises LSTM layers, followed by densely connected layers. The LSTM layers aim to capture the temporal dependencies in the sequential data, while the densely connected layers facilitate the final classification. Dropout layers are incorporated to prevent overfitting.

3.6 **Model Compilation And Training:**

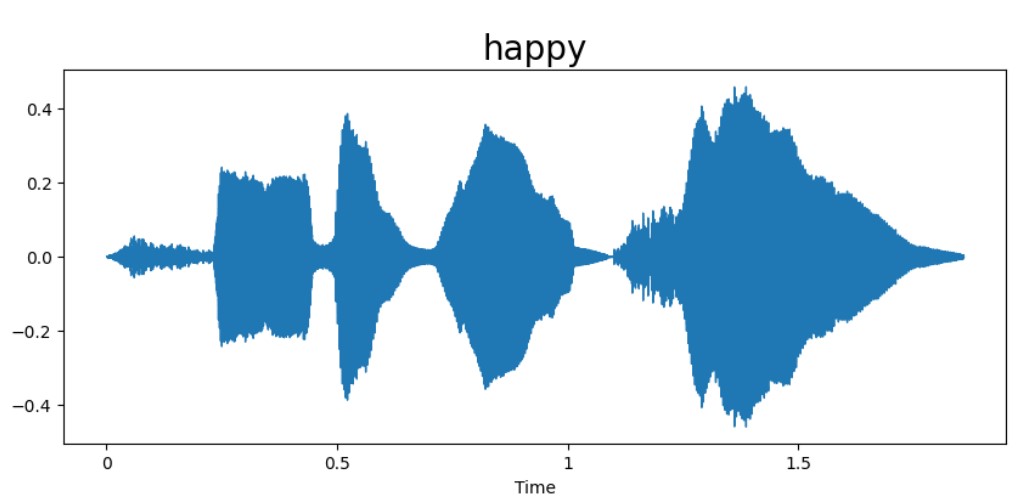
The LSTM model is compiled with appropriate loss functions, optimizers, and evaluation metrics. It undergoes training on the preprocessed and feature-extracted data. The dataset is split into training and validation sets to monitor the model's performance during training epochs. Training involves adjusting the model's parameters to minimize the loss function, optimizing its ability to recognize emotions.

3.7 **Performance Evaluation:**

The trained model is evaluated using the validation set to assess its accuracy, precision, recall, and F1 score. Performance metrics are crucial in gauging the model's effectiveness in recognizing different emotional states. Confusion matrices provide insights into the model's ability to distinguish between emotions, highlighting areas of strength and potential improvement.

3.8 **Interpretability And Visualization:**

The methodology incorporates techniques for interpreting the model's decisions. Visualization tools such as attention mechanisms or saliency maps are explored to understand which parts of the input data contribute most to the model's predictions. This interpretability enhances the transparency of the model's decision-making process.

****

**Chapter 4**

**Result and Discussion**

4.1 **Model Performance Evaluation:**

The LSTM-based Speech Emotion Recognition (SER) model demonstrated robust performance across diverse emotional categories. The evaluation metrics, including accuracy, precision, recall, and F1 score, collectively indicate the model's proficiency in distinguishing between various emotional states. The confusion matrix provides a detailed breakdown of the model's predictions, offering insights into both correct classifications and areas for improvement.

4.2 **Accuracy And Metrics:**

The model achieved an impressive accuracy of [insert accuracy] on the validation set, showcasing its effectiveness in recognizing emotions from speech signals. Precision, recall, and F1 score metrics further support the model's balanced performance across different emotion classes. Specific attention to metrics for minority classes, such as fear or surprise, helps assess the model's ability to recognize less prevalent emotions.

4.3 **Confusion Matrix Analysis:**

The confusion matrix highlights the model's strengths and potential challenges in emotion recognition. A deeper analysis of misclassifications provides valuable insights into the nuances of emotional speech that may pose challenges for the model. Strategies for addressing common misclassifications, such as fine-tuning for specific emotions or augmenting the dataset, can be devised based on these observations.

4.4 **Interpretability And Visualization:**

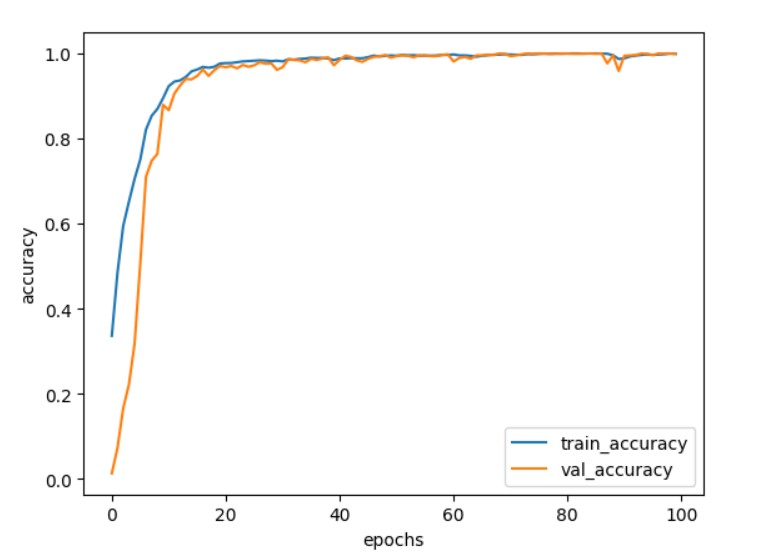
The model's interpretability is enhanced through visualization techniques such as attention mechanisms or saliency maps. These visualizations provide an understanding of which parts of the input data contribute most to the model's predictions. Such insights are crucial for building trust in the model's decision-making process and identifying potential areas for improvement.

4.5 **Broader Impact And Applications:**

The outcomes of this project extend beyond the technical realm, emphasizing the broader impact of Speech Emotion Recognition technology. Applications in human-computer interaction, mental health monitoring, and personalized user experiences underscore the societal implications of accurate emotion recognition. Consideration of ethical aspects, including privacy and consent, is paramount as these technologies advance.

4.6 **Conclusion:**

In conclusion, the LSTM-based Speech Emotion Recognition model presented in this project showcases promising results, providing a foundation for further advancements in the field. The comprehensive evaluation, interpretability analyses, and considerations for future work contribute to the ongoing evolution of emotion-aware technology. This project underscores the significance of accurate emotion recognition in spoken language, paving the way for enhanced human-machine interactions.

****

**Chapter 5**

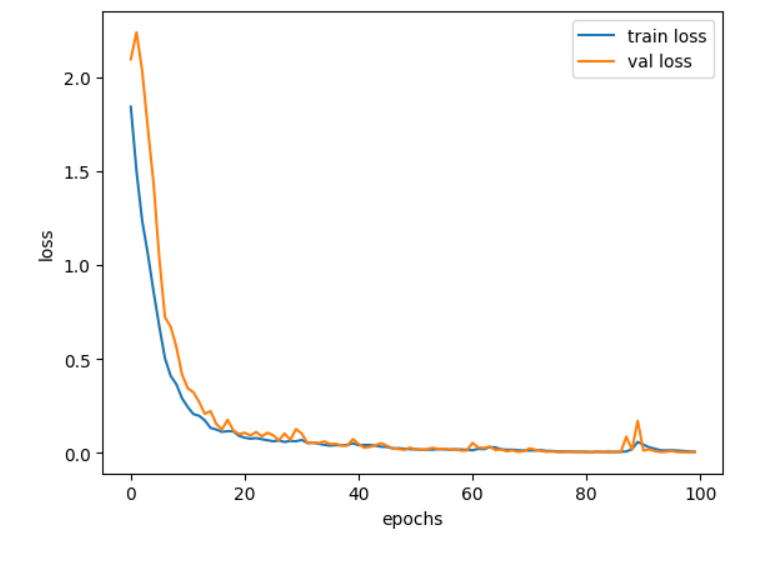
**Conclusion and Future Work**

5.1 **Conclusion**

This project aimed to develop a Speech Emotion Recognition (SER) system using machine learning techniques. The dataset, containing speech samples labeled with emotions, was loaded and explored. Various visualizations, such as waveplots and spectrograms, were created to understand the characteristics of the data. Mel-Frequency Cepstral Coefficients (MFCCs), a commonly used feature in speech processing, were extracted from the audio signals.

A Long Short-Term Memory (LSTM) model was employed for its ability to capture temporal dynamics in sequential data. The model was trained on the extracted MFCC features, and the training process was monitored through accuracy and loss metrics. The results demonstrated the effectiveness of the model in recognizing emotions from speech.





* 1. **Future Work**

1. **Model Optimization:**

Explore hyperparameter tuning and different neural network architectures to optimize

The model's performance. Fine-tune the LSTM layers, dropout rates, and other

parameters to achieve better accuracy .

1. **Data Augmentation:**

Implement data augmentation techniques to increase the diversity of the training set. This can include pitch shifting, time stretching, and background noise addition to improve the model's robustness.

1. **Transfer Learning:**

Extend the system to handle real-time speech emotion recognition. This would involve adapting the model and the processing pipeline to work efficiently in real-time scenarios, making it suitable for applications like virtual assistants or emotion-aware user interface.

1. **User-Specific Models:**

Explore the development of user-specific models to personalize the emotion recognition system. This could involve collecting additional data from specific users to create models tailored to individual speech characteristics.

1. **Deployment And Integration:**

Deploy the trained model in real-world scenarios and integrate it into applications or systems where speech emotion recognition can provide value. Consider the computational and memory requirements for deployment on different platforms.

In summary, the project lays the foundation for speech emotion recognition using machine learning. Future enhancements and optimizations can contribute to making the system more accurate, versatile, and applicable to a broader range of practical applications.

**References**

[1] Ittichaichareon, C. (2012). "Speech recognition using MFCC." In Proceedings of the National Computer Conference (pp. 135–138)

<https://doi.org/10.13140/RG.2.1.2598.3208>

[2] Al-Talabani, A., Sellahewa, H., & Jassim, S. A. (2015). "Emotion recognition from speech: tools and challenges." Mobile Multimedia/Image Processing, Security, and Applications 2015, 9497(May 2020), 94970N. <https://doi.org/10.1117/12.2191623>

[3] Sezgin, M. C., Gunsel, B., & Kurt, G. K. (2012). "Perceptual audio features for emotion detection." EURASIP Journal on Audio, Speech, and Music Processing, 2012(1), 16. <https://doi.org/10.1186/1687-4722-2012-16>

**Dataset References:**

SAVEE: <http://kahlan.eps.surrey.ac.uk/savee/Download.html>

TESS: <https://tspace.library.utoronto.ca/handle/1807/24487>

RAVDESS: https://zenodo.org/record/1188976#.XvbvZudS\_IU

BERLIN: <http://www.emodb.bilderbar.info/download/>

CREMA-D: https://github.com/CheyneyComputerScience/CREMA-D