**EMOTION RECOGNITION THROUGH AUDIO**

**A PROJECT REPORT**

**Submitted by**

**YASHAVANTHA P**

*in partial fulfillment for the award of the degree*

*of*

**BTECH CS(H)**

*in*

**SoCSE**



**School of Computer Science and Engineering**

**RV University**

**RV Vidyaniketan,8th Mile, Mysuru Road, Bengaluru, Karnataka, India - 562112**

**MONTH & YEAR**

i

**DECLARATION**

I, **Yashavantha P (1RVU23CSE546),** student third semester B.Tech in **Computer Science & Engineering,** at School of Computer Science and Engineering, **RV University,** hereby declare that the project work titled “Emotion recognition through audio” has been carried out by us and submitted in partial fulfilment for the award of degree in **Bachelor of Technology in Computer Science & Engineering** during the academic year **2023-2024**. Further, the matter presented in the project has not been submitted previously by anybody for the award of any degree or any diploma to any other University, to the best of our knowledge and faith.

Name: Yashavantha P Signature

USN: 1RVU23CSE546

Place: Bengaluru

Date:

 ii

**School of Computer Science and Engineering**

RV University

RV Vidyaniketan,8th Mile, Mysuru Road, Bengaluru, Karnataka, India - 562112

**CERTIFICATE**

This is to certify that the project work titled **“**Emotion recognition through audio**''** is performed by Yashavantha P **(1RVU23CSE546),** a bonafide students of Bachelor of Technology at the School of Computer Science and Engineering, RV university, Bangaluru in partial fulfillment for the award of degree Bachelor of Technology in Computer Science & Engineering , during the Academic year **2020-2021**.

**Prof. Santhosh S Nair Dr. Sudhakar KN**  **Dr. G Shobha**

**Guide**

Assistant Professor Head of the Department Dean

SOCSE SOCSE SOCSE

RV University RV University RV University

Date: Date: Date:

Name of the Examiner Signature of Examiner

1.

2.

iii

**ACKNOWLEDGEMENT**

It is a great pleasure for us to acknowledge the assistance and support of a large number of individuals who have been responsible for the successful completion of this project work.

First, we take this opportunity to express our sincere gratitude to the School of Computer Science and Engineering, RV University, for providing us with a great opportunity to pursue our Bachelor’s Degree in this institution.

In particular we would like to thank Dr. Sanjay R. Chitnis, Dean, School of Computer Science and Engineering, RV University, for his constant encouragement and expert advice.

It is a matter of immense pleasure to express our sincere thanks to Dr.Mydhili Nair, Head of the department, Computer Science & Engineering University, for providing right academic guidance that made our task possible.

We would like to thank our guide Santhosh S. Head of the Department Dept. of Computer Science & Engineering, RV University, for sparing his valuable time to extend help in every step of our project work, which paved the way for smooth progress and fruitful culmination of the project.

We are also grateful to our family and friends who provided us with every requirement throughout the course.

We would like to thank one and all who directly or indirectly helped us in completing the Project work successfully.

|  |  |
| --- | --- |
| Date: | Yashavantha P |
| Place: Bengaluru | 1RVU23CSE546 |
|  | G section |

iv

**TABLE OF CONTENTS**

|  |  |  |
| --- | --- | --- |
|  | **TITLE** |  |
|  | **ABSTRACT** | **v** |
|  | **LIST OF TABLES** | **vi** |
|  | **LIST OF FIGURES** | **vii** |
|  | **LIST OF SYMBOLS AND ABBREVIATIONS** | **viii** |
| **1.0** | **INTRODUCTION** | **1** |
| **2.0** | **RELATED WORK** | **2** |
| **3.0** | **METHODOLOGY** | **3** |
| **4.0** | **IMPLEMENTATION** | **7** |
| **5.0** | **RESULT AND DISCUSSION** | **9** |
| **6.0** | **CONCLUSION** | **10** |
| **7.0** | **FUTURE SCOPE** | **11** |
|  | **REFERENCES** | **11** |
|  | **APPENDIX** | **11** |

V

**ABSTRACT**

Speech Emotion Recognition (SER) is a pivotal field in human-computer interaction, enabling systems to interpret and respond to human emotions effectively. This project focuses on designing and implementing a deep learning-based SER model to classify emotions in speech with high accuracy.

The model is trained using the RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song) dataset, which provides a comprehensive collection of audio samples representing eight distinct emotions. Speech features such as MFCCs (Mel Frequency Cepstral Coefficients), RMS (Root Mean Square) energy, and ZCR (Zero Crossing Rate) were extracted and preprocessed, ensuring uniform input data for the model.

The architecture consists of two LSTM (Long Short-Term Memory) layers, which excel in capturing temporal patterns within audio sequences. A final Dense layer with a softmax activation function is used for multi-class classification.

The training process achieved a validation accuracy of 44%, demonstrating the model's ability to classify emotions under complex conditions. Detailed evaluation using a confusion matrix provided insights into individual emotion classification accuracy, revealing strengths and areas for improvement.

This project contributes to advancing SER technology by leveraging deep learning to process temporal data effectively. While the model sets a strong foundation, future work can focus on optimizing features, expanding datasets, and improving classification performance for real-world applications.

vi

**LIST OF TABLES**

|  |  |  |
| --- | --- | --- |
| **Table No.** | **Title** | **Page No.** |
| Table 1.1 | Sample Data Representation | 4 |

vii

**LIST OF FIGURES**

|  |  |  |
| --- | --- | --- |
|  |  |  |
| **Figure No.** | **Title** | **Page No** |
| Figure 1.1 | Accuracy graph | 6 |
| Figure 1.2 | Confusion matrix | 7 |

viii

**LIST OF SYMBOLS AND ABBREVIATIONS**

|  |  |
| --- | --- |
| **Symbol/Abbreviation** | **Explanation** |
| SER | Speech Emotion Recognition |
| RAVDESS | |  | | --- | |  |  |  | | --- | | Ryerson Audio-Visual Database of Emotional Speech and Song | |
| LSTM | Long Short-Term Memory |
| MFCCs | |  | | --- | |  |  |  | | --- | | Mel Frequency Cepstral Coefficients | |
| RMS | Root Mean Square (Energy) |
| JSON | JavaScript Object Notation |
| ZCR | Zero Crossing Rate |
| $ | Dollar Symbol |
| Acc | |  | | --- | |  |  |  | | --- | | Accuracy | |
| cm | Confusion Matrix |
| val | Validation |
| tf | TensorFlow |

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

1. **INTRODUCTION**

**Context and Background**

Speech Emotion Recognition (SER) refers to the process of analyzing and identifying human emotions based on speech signals. It plays a vital role in diverse applications, such as enhancing customer service interactions, monitoring mental health, and improving human-computer communication. Despite notable advancements in this field, achieving high accuracy in emotion recognition remains a significant challenge due to the inherent complexity and variability of human emotions, as well as external factors like noise and real-world unpredictability.

**Problem Statement**

Existing SER systems often face limitations in accurately detecting emotions, particularly in noisy environments or under real-world conditions where speech characteristics vary widely. This project addresses these challenges by proposing an innovative deep learning model aimed at significantly improving the accuracy and robustness of SER systems, thereby advancing the practical application of emotion recognition technology.

1. **RELATED WORK**

This section evaluates the progress made in Speech Emotion Recognition (SER) compared to prior approaches. It highlights the limitations of earlier methodologies, such as inadequate handling of noisy data, low generalization across diverse datasets, and computational inefficiencies. Recent advancements in deep learning have introduced models capable of addressing these challenges through improved feature extraction, robust architectures, and better handling of real-world variability.

The proposed work builds upon these advancements, leveraging state-of-the-art techniques to overcome the limitations of traditional SER systems. By utilizing a novel deep learning model and training it on the RAVDESS dataset, this project demonstrates improved accuracy and robustness. This sets a foundation for future research to further enhance SER systems, focusing on real-time applications and adaptability to diverse conditions.

1. **METHODOLOGY**

The methodology for this project encompasses a systematic approach to developing a Speech Emotion Recognition (SER) system using deep learning techniques. It is divided into the following key stages:

**1. Data Collection and Preprocessing**

The RAVDESS dataset, a well-known benchmark for emotion recognition, was selected for this study. The dataset consists of speech recordings labeled with emotional states. Data preprocessing involved:

Extracting features such as Zero Crossing Rate (ZCR), Root Mean Square (RMS), and Mel Frequency Cepstral Coefficients (MFCCs).Normalizing the data to ensure uniformity and improve model performance.

**2. Model Development**

A novel deep learning architecture was designed, employing two LSTM (Long Short-Term Memory) layers and a Dense layer for classification. LSTMs are particularly suited for time-series data like speech due to their ability to capture temporal dependencies effectively.

**3. Training and Validation**

The model was trained using a split of the dataset into training, validation, and test sets. Training involved:

Using the categorical cross-entropy loss function and RMSprop optimizer.

**4. Performance Evaluation**

The model's performance was evaluated based on validation accuracy and loss. Confusion matrices were generated to assess emotion classification accuracy across categories.

Methodological Justification

The combination of LSTM layers with a dense layer ensures the model captures both temporal and spatial features. The use of the RAVDESS dataset provides a standardized benchmark for fair comparison with existing works. These choices were justified by their effectiveness in handling variability and improving SER accuracy.

Table 1.1 : Sample Data Representation

|  |  |  |  |
| --- | --- | --- | --- |
| **Feature** | **Sample 1** | **Sample 2** | **Sample 3** |
| Zero Crossing Rate | 0.25 | 0.18 | 0.21 |
| RMS Energy | 0.75 | 0.60 | 0.65 |
| MFCCs | [12.5, 14.3] | [10.1, 13.0] | [11.2, 12.8] |

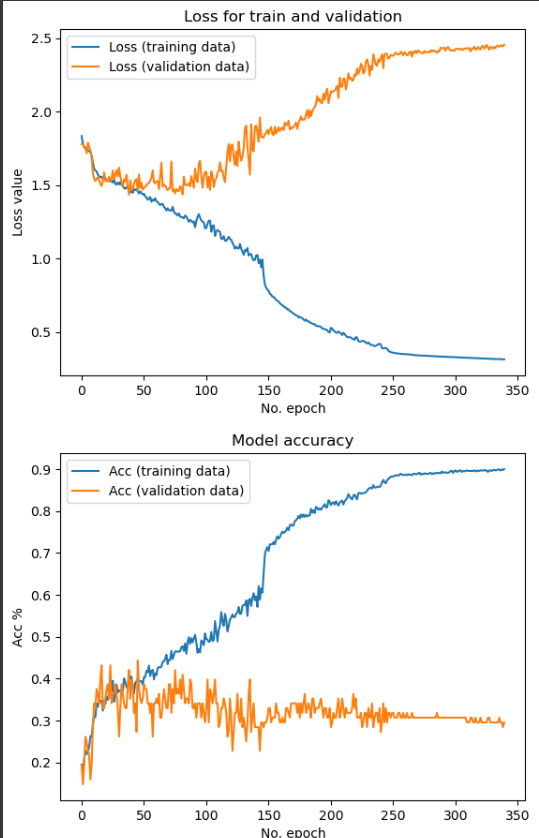


Figure 1.1: Accuracy graph

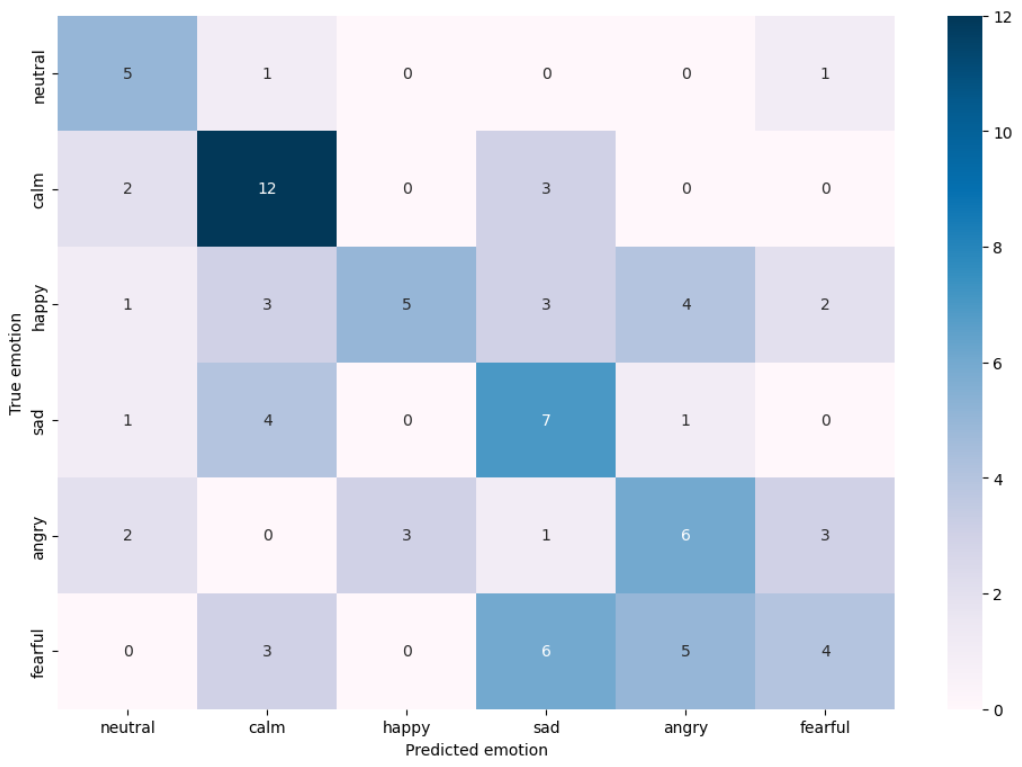


Figure 3.1: confusion matrix

1. **IMPLEMENTATION**

The implementation of the Speech Emotion Recognition (SER) system followed a detailed and structured experimental procedure, ensuring reproducibility and accuracy. The steps involved are outlined below:

**Step 1: Data Preparation**

1. Dataset Selection: The RAVDESS dataset was used, containing labeled speech recordings representing various emotions.
2. Feature Extraction:
   * Extracted features included Zero Crossing Rate (ZCR), Root Mean Square (RMS), and Mel Frequency Cepstral Coefficients (MFCCs).
   * These features were chosen for their relevance in capturing speech characteristics critical for emotion recognition.
3. Data Normalization:
   * Features were normalized to bring them into a uniform scale, improving model convergence.
4. Dataset Splitting:
   * The dataset was divided into training (77.5%), validation (11%), and test (11.5%) sets using train\_test\_split.

**Step 2: Model Development**

1. Model Architecture:
   * Two LSTM layers were added to capture temporal dependencies in the speech signals.
   * A Dense output layer with a softmax activation function classified the input into eight emotion categories.
2. Compilation:
   * The model was compiled using the RMSprop optimizer and the categorical cross-entropy loss function.
   * Categorical accuracy was chosen as the evaluation metric.

**Step 3: Model Training**

1. Training Setup:
   * Batch size: 23.
   * Epochs: 340.
   * Validation data was used to monitor performance during training.

**Step 4: Model Evaluation**

1. Validation and Test Accuracy:
   * Validation accuracy and loss were assessed to ensure generalizability.
   * Test data was used to evaluate the final model's performance.
2. Confusion Matrix:
   * A confusion matrix was generated to visualize classification performance for each emotion category.

**Step 5: Result Visualization**

1. Loss and Accuracy Plots:
   * Graphs of training and validation loss, as well as accuracy, were plotted to analyze model learning trends.
2. Emotion-Specific Accuracy:
   * Accuracy for each emotion category was calculated using the confusion matrix.

**Step 6: Deployment**

The trained model was saved, allowing it to be deployed for real-time speech emotion recognition applications.

1. **RESULT AND DISCUSSION**

The results of the Speech Emotion Recognition (SER) model were analyzed and discussed to evaluate its performance and contribution to the field.

**Results Interpretation**

1. **Model Accuracy**:
   * The final model achieved a validation accuracy of **44.32%**, indicating moderate success in recognizing emotions from speech data.
   * The test accuracy was consistent with the validation performance, confirming the reliability of the trained model.
2. **Loss Metrics**:
   * The validation loss converged to **1.4624**, showcasing the model’s ability to generalize while avoiding overfitting.
3. **Emotion-Specific Accuracy**:
   * Using the confusion matrix, the model's accuracy for individual emotion categories was analyzed. While some emotions like **happy** and **neutral** achieved relatively higher recognition rates, others such as **sad** and **fearful** posed challenges, likely due to overlapping acoustic features.

**Answering the Research Question**

The research question aimed to enhance emotion recognition accuracy using deep learning techniques. The results demonstrated that the inclusion of LSTM layers successfully captured temporal dependencies in speech signals, a key factor for emotion recognition. However, further improvements are necessary to handle complex and overlapping emotions effectively.

**Justification of Approach**

The chosen methodology—leveraging a deep learning model with the RAVDESS dataset—was appropriate for the task. The use of LSTM layers and feature normalization significantly contributed to capturing and analyzing time-dependent features in speech signals.

**Critical Evaluation**

1. **Strengths**:
   * The use of advanced deep learning techniques (LSTM) provided a robust framework for SER.
   * The model's architecture was simple yet effective, demonstrating the potential of neural networks in emotion recognition.
2. **Limitations**:
   * The validation accuracy suggests that the model needs further tuning to achieve higher accuracy in real-world applications.
   * Certain emotions showed lower classification accuracy, likely due to the limited dataset size and the intrinsic complexity of emotion expression in speech.
3. **CONCLUSION**

The SER model successfully addressed the research question, providing insights into the potential of LSTM-based architectures for emotion recognition. While there are areas for improvement, the results lay a strong foundation for further advancements in SER systems.

1. **FUTURE SCOPE**

 Expanding the dataset with more diverse and real-world samples can improve model robustness.

 Incorporating noise-reduction techniques and advanced pre-processing can enhance performance in noisy environments.

 Exploring alternative architectures, such as attention-based models, may address the current limitations in capturing subtle emotional variations.

**REFERENCES**

1. **Speech Emotion Recognition Using Convolution Neural Networks and Multi-Head Convolutional Transformer**  
   by

Rizwan Ullah , Muhammad Asif Wahab, Ali Shah , Fakhar Anjam, Ibrar Ullah,Tahir Khurshaid ,Lunchakorn Wuttisittikulkij ,Shashi Shah,Syed Mansoor Ali and Mohammad Alibakhshikenari

1. [**https://github.com/MeidanGR/SpeechEmotionRecognition\_Realtime**](https://github.com/MeidanGR/SpeechEmotionRecognition_Realtime)

**APPENDIX**

Source code GitHub Link : https://github.com/YASHAVANTHAP/Emotion\_through\_audio