Detection of Emotion of Speech using Hybrid CNN

A

Project Report

submitted in partial fulfillment of the

requirements for the award of the degree of

BACHELOR OF TECHNOLOGY

in

COMPUTER SCIENCE & ENGINEERING

by

Name	Roll No.
Vartika Rawat	R2142201670
Komal Rajpoot	R2142201635
Satyam Raj	R2142201917
Vansh Gupta	R2142201652

under the guidance of

Dr. Bhupendra Singh

School of Computer Science

University of Petroleum & Energy Studies

Bidholi, Via Prem Nagar, Dehradun, Uttarakhand

December – 2023

CANDIDATE'S DECLARATION

We hereby certify that the project work entitled **"Detection of Emotion of Speech using Hybrid CNN"** in partial fulfillment of the requirements for the award of the Degree of BACHELOR OF TECHNOLOGY in COMPUTER SCIENCE AND ENGINEERING with specialization in Artificial Intelligence and Machine Learning) and submitted to the Department of Informatics, School of Computer Science, University of Petroleum & Energy Studies, Dehradun, is an authentic record of our work carried out during a period from **January**, **2024** to **May**, **2024** under the supervision of **Dr. Bhupendra Singh** ***(Assistant Professor (SS)).***

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

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

Date: 03/ March/ 2024

(Dr. Bhupendra Singh)
Project Guide

ACKNOWLEDGEMENT

We wish to express our deep gratitude to our guide **Dr. Bhupendra Singh**, for all advice, encouragement and constant support he has given us throughout our project work. This work would not have been possible without his support and valuable suggestions.

We sincerely thanks to our respected **Dr. Anil Kumar, Head Department of Artificial Intelligence**, for his great support in doing our project in **Detection of Emotion of Speech using Hybrid CNN.**

We are also grateful to Dean SoCS UPES for giving us the necessary facilities to carry out our project work successfully. We also thanks to our Course Coordinator, (Dr. Mohammad Ahsan) and our Activity Coordinator (Dr. Mohammad Ahsan) for providing timely support and information during the completion of this project.

We would like to thank all our **friends** for their help and constructive criticism during our project work. Finally, we have no words to express our sincere gratitude to our **parents** who have shown us this world and for every support they have given us.

Name	**Vartika Rawat**	**Komal Rajpoot**	**Satyam Raj**	**Vansh Gupta**
Roll No.	**R2142201670**	**R2142201635**	**R2142201917**	**R2142201652**

ABSTRACT

Emotion detection from speech signals is a crucial task with numerous applications in human-computer interaction, sentiment analysis, and affective computing. In this paper, we propose a novel approach for emotion recognition from speech by employing a hybrid Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) model. We utilize the RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song) dataset, which contains acted speech from actors portraying various emotional states.

First, we preprocess the speech data to extract relevant features such as Mel-frequency cepstral coefficients (MFCCs) and spectrograms. These features capture both temporal and spectral characteristics of the speech signals. We then construct a CNN-LSTM architecture to learn hierarchical features from the input data.

The CNN component extracts local features from the spectrograms, capturing patterns and structures within short-time frames. Subsequently, the LSTM component processes these features over time, capturing long-term dependencies and temporal dynamics in the speech signals. The hybrid architecture allows our model to effectively capture both local and temporal information, leading to improved emotion recognition performance.

TABLE OF CONTENTS

INTRODUCTION

1.1 History

The ability to understand human emotions through speech has garnered significant attention. This capability not only enhances various technological systems but also opens avenues for improving human experiences across diverse fields. From education to healthcare, from automotive to security, the potential applications of emotion recognition technology are vast and promising.

Emotion recognition systems analyze speech to discern underlying emotions accurately. These systems find applications in education, where they can adapt learning materials based on students' emotional states, thereby enhancing engagement and performance. In the automotive industry, such systems can enhance driving experiences by understanding drivers' emotions and adjusting accordingly. In security, they contribute to public safety by detecting extreme emotions like fear and anxiety in crowded spaces. Additionally, in communication settings like call centers, integrating emotion recognition into customer service processes can lead to more personalized interactions and improved satisfaction.

While the focus primarily remains on analyzing visual and auditory signals, particularly acoustic signals, advancements in machine learning have introduced various models for emotion recognition. From traditional approaches like K Nearest Neighbors and decision trees to deep learning techniques such as Deep Neural Networks (DNNs) and Convolutional Neural Networks (CNNs), the landscape of emotion recognition technology continues to evolve.

1.2 Requirement Analysis

2. **Data Collection:** Gather a dataset of speech samples labeled with different emotions (e.g., happy, sad, angry, etc.) in our case it is RAVDESS speech dataset.
3. **Feature Extraction:** Extract relevant features from the audio data, such as MFCCs (Mel-frequency cepstral coefficients), spectrograms, or other time-frequency representations.
4. **Pre-processing:** Clean the audio data, remove noise, and possibly segment it into smaller chunks for analysis.
5. **Model Architecture Design**: Design a hybrid CNN architecture that combines convolution layers with other types of layers (e.g., recurrent layers, dense layers in our case LSTM) to effectively learn features from the audio data.
6. **Model Training:** Train the hybrid CNN model on the pre-processed and feature-extracted data, using appropriate training techniques such as data augmentation to prevent over fitting.
7. **Model Evaluation**: Evaluate the trained model using metrics such as accuracy, precision, recall, and F1-score on a separate validation set to assess its performance.
8. **Testing**: Test the final model on unseen data to assess its generalization ability and real-world performance.
9. **Analysis and Interpretation**: Analyze the results, interpret the model's behavior, and identify areas for improvement or further research.

1.3 Main Objective

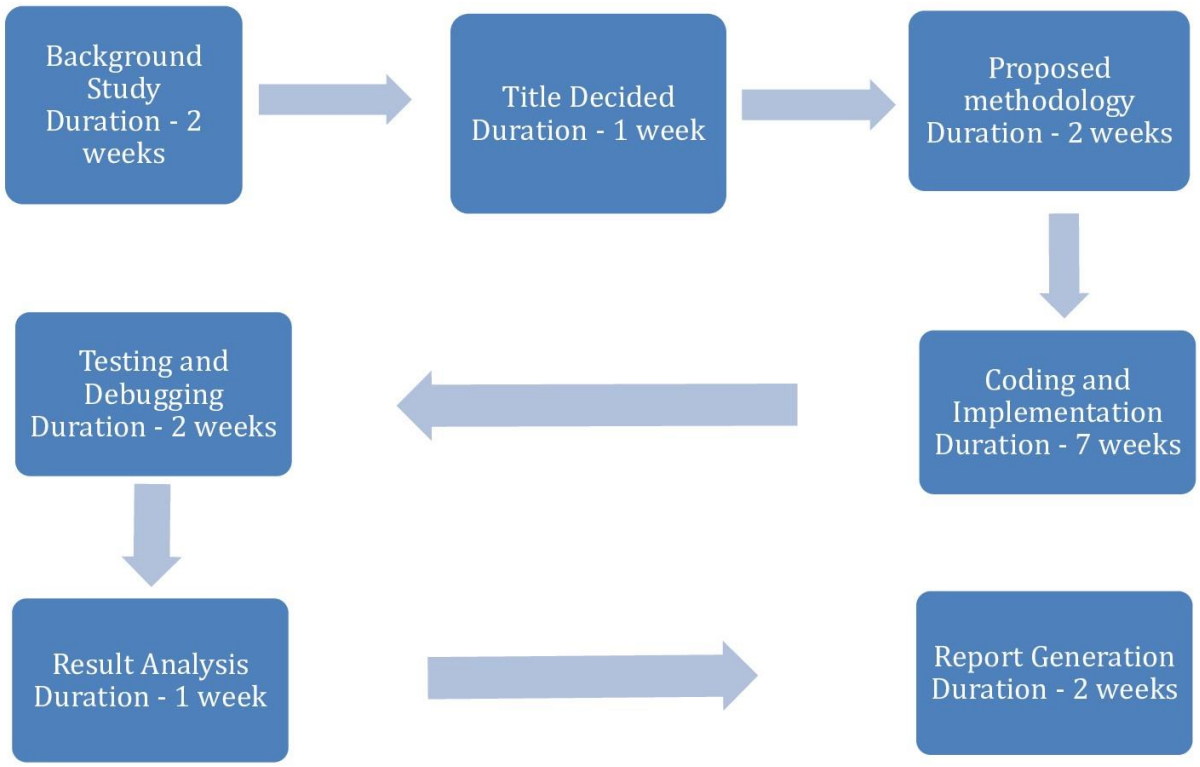
The necessity for Speech Emotion Recognition (SER) systems to complement existing automated systems. The need for accurate and efficient SER models capable of analyzing speech signals and detecting emotional states.

1.4 Sub Objectives

The challenge of selecting suitable speech segments and extracting relevant features to improve SER performance.

The ongoing quest to optimize SER algorithms and classification techniques to achieve higher recognition rates and enhance user experience.

1.5 Pert Chart Legend



2 SYSTEM ANALYSIS

2.1 Existing System

- **Data Collection and Preprocessing**

  The existing system utilizes the RAVDESS dataset, which contains audio recordings of actors performing various emotional states. Preprocessing techniques such as audio normalization, feature extraction (e.g., MFCCs), and data augmentation may be applied to enhance the quality and diversity of the training data.

- **Hybrid CNN Architecture:**

  The existing system employs a hybrid CNN architecture that combines both 1D and 2D convolutional layers to capture temporal and spatial features from the audio input. The 1D convolutional layers extract temporal features from the audio signals, while the 2D convolutional layers capture spatial patterns from spectrogram representations.

- **Model Training:**

  The hybrid CNN model is trained on the preprocessed RAVDESS dataset using techniques such as mini-batch stochastic gradient descent (SGD) or Adam optimizer. During training, the model learns to classify emotional states (e.g., happiness, sadness, anger) based on the features extracted from the audio input.

- **Evaluation and Validation:**

  The trained model is evaluated using standard evaluation metrics such as accuracy, precision, recall, and F1-score on a separate validation dataset. Cross-validation techniques may be employed to assess the generalization performance of the model and identify potential overfitting.

- **Performance Analysis**

  The performance of the existing system is analyzed in terms of its ability to accurately classify emotional states from speech signals. Confusion matrices and receiver operating characteristic (ROC) curves may be used to visualize the model's performance across different emotion classes.

- **Comparison with Baseline Models:**

  The existing system may be compared with baseline models such as traditional machine learning classifiers or single-modality CNNs to demonstrate its superiority in terms of accuracy and robustness.

- **Limitations and Future Work:**

  The existing system may have limitations such as computational complexity, sensitivity to noise, or difficulties in capturing subtle emotional cues. Future work may focus on addressing these limitations by exploring advanced model architectures, incorporating multimodal information, or leveraging transfer learning techniques.

- **Real-World Applications:**

  The existing system has applications in various domains including affective computing, human-computer interaction, virtual assistants, mental health monitoring, and market research.

2.2 Motivations

As we know, speech is a direct way to transfer information from one person to another. It contains a wide variety of information and expresses rich emotional information through the emotions it contains. The objective of this project is to develop a hybrid convolutional neural network (CNN) model capable of accurately identifying and categorizing emotions expressed in speech signals. The goal is to utilize the features extracted by the CNN architecture to improve the classification performance for emotion detection tasks in speech data.

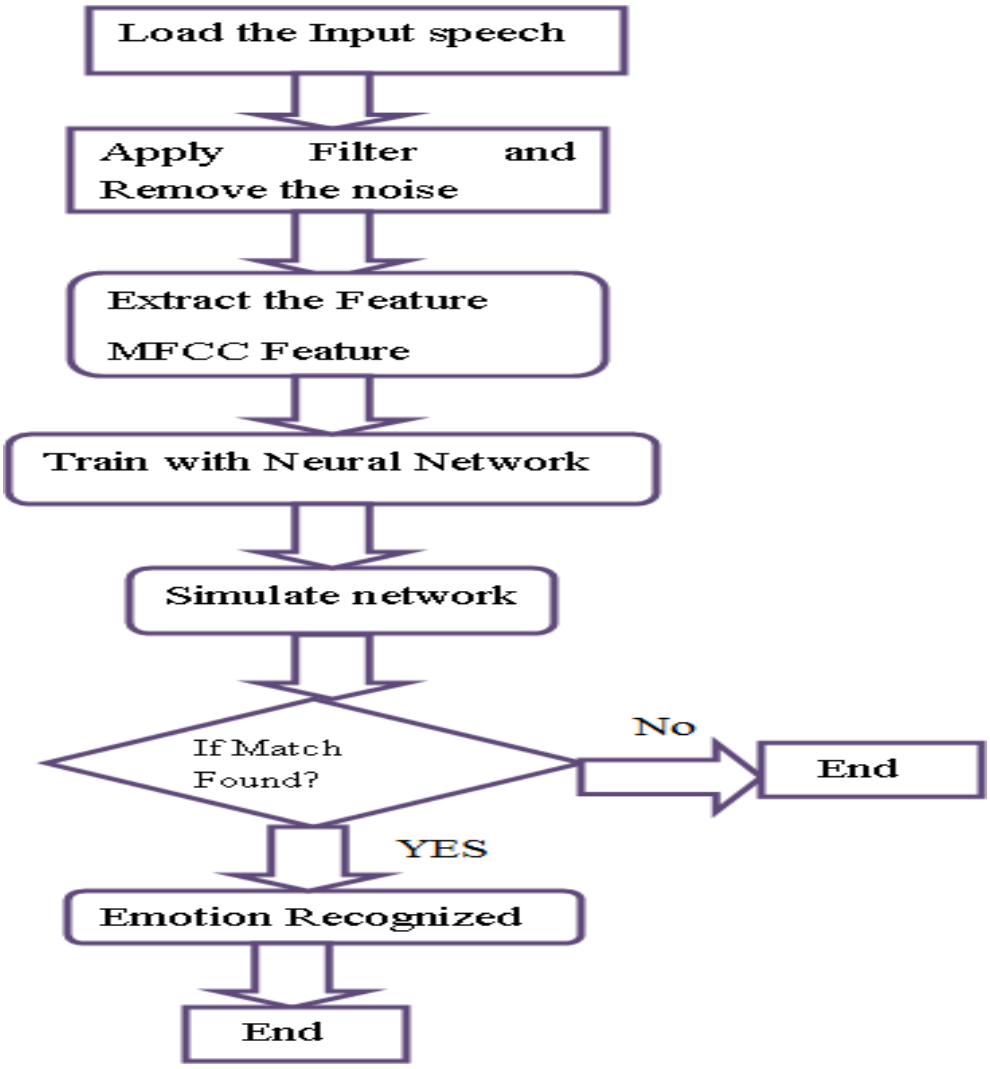
2.3 Proposed System

The proposed system for Detection of Emotion of Speech using Hybrid CNN . It involves the development of a model, likely based on neural networks that can analyze relevant data to accurately identify the presence or progression of Emotions. Preprocessing Audio data preprocessing involves converting raw audio signals into spectrograms or other suitable representations. Features extraction from audio signals, including MFCCs, spectral features, and prosodic features. Normalization of features to ensure consistency and enhance model convergence. The proposed hybrid CNN architecture combines both 1D and 2D convolutional layers to capture both temporal and spatial features from the speech signals effectively.

2.4 Modules

- **Data Collection:** Gather a dataset of speech samples labeled with different emotions (e.g., happy, sad, angry, etc.).
- **Feature Extraction:** Extract relevant features from the audio data, such as MFCCs (Mel-frequency cepstral coefficients), spectrograms, or other time-frequency representations.
- **Pre-processing:** Clean the audio data, remove noise, and possibly segment it into smaller chunks for analysis.
- **Model Architecture Design**: Design a hybrid CNN architecture that combines convolution layers with other types of layers (e.g., recurrent layers, dense layers) to effectively learn features from the audio data.
- **Model Training:** Train the hybrid CNN model on the pre-processed and feature-extracted data, using appropriate training techniques such as data augmentation to prevent over fitting.
- **Model Evaluation**: Evaluate the trained model using metrics such as accuracy, precision, recall, and F1-score on a separate validation set to assess its performance.
- **Testing**: Test the final model on unseen data to assess its generalization ability and real-world performance.
- **Analysis and Interpretation**: Analyze the results, interpret the model's behavior, and identify areas for improvement or further research.

3 Implementation



3.3 Scenarios

- **Call Center Emotion Analysis:**

  Scenario: A call center wants to analyze customer-agent interactions to assess customer satisfaction levels.

  Implementation: Deploy the hybrid CNN model to analyze audio recordings of customer calls in real-time. The model detects emotions expressed by both customers and agents, providing insights into the overall sentiment of the conversation.

- **Virtual Assistant Emotion Recognition:**

  Scenario: A virtual assistant application aims to respond empathetically to user queries and commands.

  Implementation: Integrate the hybrid CNN model into the virtual assistant's backend to analyze user voice commands. The model detects the user's emotional state and tailors its responses accordingly, providing a more personalized and empathetic user experience.

- **Educational Platform for Emotion-aware Learning:**

  Scenario: An online educational platform wants to enhance student engagement and motivation by providing emotion-aware feedback.

  Implementation: Integrate the hybrid CNN model into the platform's speech recognition system to analyze students' verbal responses during interactive lessons. The model detects emotions such as confusion or frustration and provides targeted feedback and support to improve learning outcomes.

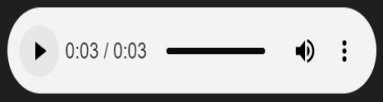
3.2 Algorithms

- CNN architecture is designed to effectively capture spatial features from spectrogram representations of speech signals, which are then used for emotion detection in the hybrid CNN model.

- In the project "Detection of Emotion of Speech Using Hybrid CNN," the Long Short-Term Memory (LSTM) network is employed to capture temporal dependencies in the sequential data, which is essential for understanding the emotional content conveyed through speech signals.By capturing temporal dependencies in the sequential data, LSTM enhances the model's ability to understand the emotional content conveyed through speech, leading to more accurate emotion detection.

4 Output screens

```
[60]
# Assuming the file path is available
wav_file_path = '/content/drive/MyDrive/ravdess/speech/Actor_01/03-01-01-01-01-01-01.wav'  # Update with the actual path
Audio(wav_file_path)
```




```
[61]
predict('/content/drive/MyDrive/ravdess/speech/Actor_01/03-01-01-01-01-01-01.wav')
```

```
1/1 [==============================] - 0s 100ms/step
neutral
```

```
[33]
cnn_lstm_model.summary()
```

```
Model: "sequential_3"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 conv1d_3 (Conv1D)           (None, 36, 64)            384

 max_pooling1d_3 (MaxPoolin  (None, 18, 64)            0
 g1D)

 lstm_3 (LSTM)               (None, 128)               98816

 dense_9 (Dense)             (None, 128)               16512

 dropout_6 (Dropout)         (None, 128)               0

 activation_9 (Activation)   (None, 128)               0

 dense_10 (Dense)            (None, 64)                8256

 dropout_7 (Dropout)         (None, 64)                0

 activation_10 (Activation)  (None, 64)                0

 dense_11 (Dense)            (None, 8)                 520

...
Total params: 124488 (486.28 KB)
Trainable params: 124488 (486.28 KB)
Non-trainable params: 0 (0.00 Byte)
```

5 Limitations

- **Limited Dataset Size:** The RAVDESS dataset, while widely used, is relatively small compared to some other datasets, which may limit the model's ability to generalize across diverse emotional expressions and speakers.
- **Imbalanced Class Distribution:** The RAVDESS dataset suffers from class imbalance, with certain emotions being underrepresented compared to others. This could lead to biased model predictions, especially for minority classes.
- **Limited Contextual Information:** Emotion detection from speech often requires contextual information such as body language, facial expressions, and situational context. However, the RAVDESS dataset only provides audio recordings, lacking other contextual cues that may influence emotional interpretation.
- **Limited Emotional Variability:** The emotional variability captured in the RAVDESS dataset may not fully represent the wide spectrum of human emotions encountered in real-world scenarios. This could limit the model's ability to generalize to diverse emotional expressions.
- **Potential Bias in Labeling:** The emotional labels assigned to the RAVDESS dataset may be subjective and prone to human bias, potentially affecting the model's training and evaluation. Inconsistent or inaccurate labeling could introduce noise into the training data.
- **Difficulty in Capturing Subtle Emotions:** Subtle emotional expressions, such as nuanced variations within a single emotion category, may be challenging for the model to capture accurately from speech alone. This could lead to misclassification or ambiguity in emotion prediction.
- **Overfitting Concerns:** Due to the relatively small size of the RAVDESS dataset, there is a risk of overfitting, particularly when training complex hybrid CNN architectures. Careful regularization techniques and model validation are necessary to mitigate this risk.
- **Difficulty in Generalization:** Despite the use of a hybrid CNN architecture, the model's ability to generalize to unseen speakers, languages, and cultural contexts may be limited. Transfer learning or domain adaptation techniques may be necessary to improve generalization performance.

6 Future Enhancements

Future enhancements in emotion detection from speech could involve the hybridization of Convolutional Neural Networks (CNN) with recurrent or attention-based models for better temporal feature extraction and context understanding. Incorporating multi-modal information like facial expressions or physiological signals can enrich the model's understanding of emotions. Additionally, integrating self-supervised learning or transfer learning techniques can enhance performance, particularly in low-resource scenarios. Regularization methods like dropout or batch normalization can mitigate over fitting, while ensemble methods can improve robustness. Finally, deploying efficient architectures to reduce computational overhead is crucial for real-time applications.

7 Conclusion

Our project on speech emotion recognition using a hybrid Convolutional Neural Network (CNN) has demonstrated promising results in accurately identifying emotions from spoken language. By leveraging the strengths of CNNs in feature extraction and incorporating hybrid techniques, we were able to create a robust model capable of analyzing audio signals and discerning emotional nuances. This approach offers a scalable solution for applications in customer service, mental health assessment, and human-computer interaction, among others. Although there are challenges to address, such as dataset limitations and real-world variability, our project lays a solid foundation for future work in this field. Further refinement and integration with advanced deep learning methods could lead to even more accurate and efficient systems for recognizing emotions in speech, ultimately contributing to more intuitive and empathetic technologies.

8 References

1. K. Han, D. Yu, and I. Tashev, “Speech emotion recognition using deep neural network and extreme learning machine,” in Proceedings of the 15th Annual Conference of the International Speech Communication Association Interspeech, pp. 223–227, Singapore, September 2014.
2. Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” Nature , vol. 521, no. 7553, pp. 436–444, 2015. K. He, X. Zhang, S. Ren, and J. Sun, “Spatial pyramid pooling in deep convolutional networks for visual recognition,” in Proceedings of the 13th European Conference on Computer Vision, pp. 346–3610, Springer, Zurich, Switzerland, September 2014.
3. . Huang, Y. Li, J. Tao, Z. Lian, M. Niu, and J. Yi, “Speech emotion recognition using semi-supervised learning with ladder networks. First asian conference on affective computing and intelligent interaction (ACII asia),” IEEE Deep Learning Approaches for Speech Emotion Recognition, vol. 289, pp. 1–5, 2018.