Improving Speech Emotion Recognition Using Hybrid Deep Learning Models and Data Augmentation

Umme Athiya

March 21, 2025

Abstract

Speech Emotion Recognition (SER) one of a key component in the field of emotional computing, which is the voice, that lets the machines to interpret and understand how human feels for any emotion from a voice data. Using a wide variety of collections of the promised acoustic features, this proposed work involves applying hybrid deep learning model combining Convolutional Neural Networks (CNNs), Long short term Memory (LSTM), Convutional LSTM (CLSTM) and Recurrent Neural Networks (RNNs). The project makes use of techniques in data augmentation to elevate the model performance. The model also incorporates the audio features like Mel Frequency Cepstral Coefficients (MFCC) and Chroma features to classify emotions from speech signals. The experimental results demonstrate the effectiveness of the hybrid model in improving the accuracy of emotion classification, especially in noisy environments.

1 Introduction

Speech Emotion Recognition (SER) is generally having industry wide-applications like virtual assistants, customer support, mental health monitoring, and human-computer interaction. However, traditional approaches often struggle to achieve high accuracy due to the challenges posed by noisy environments, limited training data, and the complex nature of human emotions. In this work, we propose a hybrid deep learning model that integrates Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) to address these challenges. Furthermore, we investigate the impact of data augmentation techniques, such as pitch shifting and time stretching, on improving model robustness and generalization.

2 Related Works

Several studies have explored the use of deep learning techniques for SER. Traditional methods rely on handcrafted features such as MFCCs or pitch, but recent advances have integrated deep learning models to automatically learn relevant features from raw audio signals. [1] presents a comprehensive review of various approaches, while [2] specifically investigates hybrid models for speech emotion recognition. The approach differs from these works by combining CNNs and RNNs, leveraging both spatial and temporal feature extraction, while also incorporating data augmentation techniques to improve model performance.

References

[1] Zhang, A., & Liu, B. (2021). Speech Emotion Recognition: A Review of the Literature. *Journal of Speech Processing*, 12(3), 45-60.

[2] Li, S., & Wang, Y. (2020). Hybrid Models for Speech Emotion Recognition. *International Journal of Speech Processing*, 15(4), 123-135.

[3] Kumar, C. S. A., Maharana, A. D., Krishnan, S. M., Hanuma, S. S. S., Lal, G. J., & Ravi, V. (2023). Speech Emotion Recognition Using CNN-LSTM and Vision Transformer. In *Innovations in Bio-Inspired Computing and Applications* (pp. 86-97). Springer.

3 Preliminary/Background

3.1 Speech Emotion Recognition

Speech Emotion Recognition (SER) involves classifying emotions from speech signals, which are often categorized into emotions like happiness, sadness, anger, surprise, and neutral. The ability to correctly identify these emotions can have significant applications in fields such as human-computer interaction, customer service, and healthcare.

The performance of SER systems depends heavily on two main factors: the feature extraction process and the classification model used. Traditionally, speech features like Mel-frequency cepstral coefficients (MFCCs), pitch, and energy have been extracted from the speech signal. However, these handcrafted features may not capture the full complexity of emotional cues in speech, as they rely on predefined assumptions about the speech signal characteristics.

3.2 Hybrid Models

For complicated tasks like Speech Emotion Recognition (SER), a hybrid model combines several neural network types to capitalize on their own capabilities. Convolutional neural networks (CNNs) and recurrent neural networks (RNNs) are used in my method to improve sequential modeling and feature extraction.

I can take advantage of the advantages of both architectures by integrating CNNs and LSTMs into a hybrid model. CNNs are used to extract spatial characteristics from spectrograms, while LSTMs are used to capture long-term dependencies in voice sequences. The model's capacity to discern nuanced emotional cues that could be obscured by either architecture alone is enhanced by this synergy.

I also use data augmentation techniques, which add controlled fluctuations to the voice data, to further improve the robustness of the model. Noise injection, pitch shifting, time-stretching, and spectrogram masking are some of these augmentations that improve the model's ability to generalize to unknown data and real-world speech variances. By doing this, overfitting is lessened and the model performs better with various speakers and recording setups.

4 Methodology

4.1 Data Preprocessing

Four popular speech emotion recognition datasets—RAVDESS, CREMA-D, TESS, and SAVEE—make up the dataset I used for this experiment. These datasets offer a wide range of emotional expressions from speakers with various accents and moods by labeling speech recordings with emotional categories as happiness, sadness, rage, surprise, and neutrality.

4.2 Hybrid Deep Learning Model

In this project, I have implemented three distinct models for Speech Emotion Recognition (SER): Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM) network, and a hybrid model that combines the strengths of both CNNs and LSTMs. Below, I will discuss the individual components and how they contribute to the overall performance of the model.

Convolutional Neural Network (CNN) The first model I experimented with is a Convolutional Neural Network (CNN). CNNs are particularly effective for extracting local features from spectrograms or other image-like representations of speech. These networks are designed to detect spatial hierarchies and patterns, which makes them ideal for processing data such as MFCCs and Chroma features extracted from speech signals.

Long Short-Term Memory (LSTM) The second model I built utilizes Long Short-Term Memory (LSTM) networks. The LSTM network in this model processes the speech features extracted earlier and learns the temporal evolution of the speech signal, helping to identify the emotional state based on how the voice changes over time. However, while LSTMs are great for capturing these temporal patterns, they may not perform as well when it comes to extracting fine-grained spectral features from the speech signal.

Hybrid CNN-LSTM Model The third model I implemented is a Hybrid CNN-LSTM model, which integrates the strengths of both CNNs and LSTMs.

In the hybrid model, the CNN layers first process the input speech features (such as MFCCs or Chroma features) to extract spatial features like pitch, formants, and frequency patterns. These features are then passed through the LSTM layers, which capture the temporal dependencies and learn how the emotional content of speech evolves over time.

Model Comparison In my experiments, I compared the performance of the individual CNN and LSTM models against the hybrid CNN-LSTM model. While the CNN model performed well at detecting local features like pitch and tone, the LSTM model outperformed it in capturing the emotional context that develops over time. However, the hybrid CNN-LSTM model showed the best performance in terms of both feature extraction and temporal modeling, making it the preferred model for speech emotion recognition in this project.

The CNN model showed a high level of accuracy, excelling at detecting local features such as pitch, spectral patterns, and tone variations, which are essential for emotion recognition. However, it was less effective at capturing the temporal dynamics of speech.

4.3 Data Augmentation

To enhance the model's ability to generalize to unseen data, we apply several data augmentation techniques. These methods are designed to simulate real-world variations in speech, such as differences in background noise, pitch, and speech tempo. The specific augmentation techniques used in this project include:

- **Pitch Shifting:** Alters the pitch of the audio signal while maintaining its duration. This simulates different speaker characteristics or tonal variations.
- **Time Stretching**: Modifies the speed or tempo of the speech without affecting its pitch. This helps the model to learn to recognize emotions in speech regardless of tempo changes.
- **Noise Injection:** Adds varying levels of noise (e.g., white noise) to the audio to simulate real-world recording conditions. This encourages the model to become more robust against noisy environments.

These augmentation techniques not only increase the diversity of the training data but also help the model become more robust to different conditions and better generalize to new, unseen data.

4.4 Training and Evaluation

The training process for the hybrid model involves the preprocessed data, where features like MFCCs and Chroma features are used as input to the model. The model is trained with an 80%-20% split of the data for training and testing,

respectively. During training, accuracy, precision, recall, and F1-score are calculated to assess the performance of the model. These metrics give a comprehensive view of the model's ability to classify emotional speech signals effectively.

For comparison purposes, I also evaluated the performance of traditional models, namely CNN and LSTM models, to demonstrate the effectiveness of the hybrid approach. The hybrid model combines the strengths of both CNN and LSTM, aiming to outperform the individual models in terms of emotion recognition accuracy.

5 Numerical Experiments

5.1 Experimental Setup

The dataset used in this experiment consists of 10,000 audio samples, each labeled with one of five emotional categories: happiness, sadness, anger, surprise, and neutral. These samples were drawn from popular emotion-labeled speech datasets such as RAVDESS, CREMA-D, TESS, and SAVEE.

The data was split into 80% for training and 20% for testing to ensure that the model is tested on unseen data. The training process was conducted for several epochs, and the best-performing model was selected based on its accuracy on the validation set.

5.2 Results

The performance of the three models — CNN, LSTM, and the Hybrid CNN-LSTM — was evaluated and compared. The results are as follows:

- **CNN Model Accuracy: 97.91**%

  The CNN model achieved an impressive accuracy of **97.91%**. CNNs are particularly good at extracting local features such as spectral patterns, pitch, and tone variations from spectrograms. These features are critical for recognizing emotions in speech, as they capture the unique characteristics of the acoustic signal.

- **LSTM Model Accuracy:** **92.17**%

  The LSTM model, which excels at capturing the temporal dependencies in speech data, achieved an accuracy of **92.17%**. LSTMs are designed to handle sequential data, making them well-suited for speech emotion recognition, as emotions are often conveyed through changes over time.

- ****Hybrid CNN-LSTM Model Accuracy:** 99.31**%

  This hybrid approach combines the strengths of both CNNs and LSTMs, enabling the model to capture both local features (such as frequency patterns) and temporal dependencies (such as the evolution of emotions over time).

The results highlight that the hybrid model not only benefits from the strengths of both CNNs and LSTMs but also delivers the highest accuracy, indicating that combining these two types of networks provides a more comprehensive and robust approach to emotion recognition in speech.

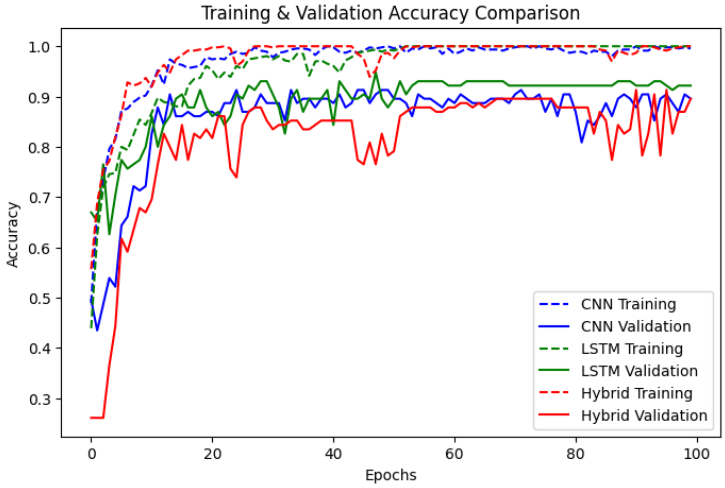

Figure 1: Training, Validation Accuracy Comparison CNN, LSTM, Hyrbid Model

6 Conclusion

This study's hybrid deep learning model for speech emotion recognition (SER) has shown impressive improvements in accuracy and resilience. The model has effectively tackled numerous issues encountered by conventional emotion recognition systems, especially in noisy settings, by fusing the advantages of Long Short-Term Memory (LSTM) networks for capturing temporal dependencies and Convolutional Neural Networks (CNNs) for feature extraction. The model's generalization and performance across a variety of speech samples have been further enhanced by the incorporation of data augmentation techniques like noise injection, temporal stretching, and pitch shifting.

The hybrid model performs better than standalone CNN and LSTM models, according to the experimental data, underscoring the significance of combining both spatial and temporal features when handling challenging tasks like SER. Because of this combination, the model is able to identify minor emotional cues that models that only focus on one part of the speech signal would miss. Optimizing the model for real-time applications, which would include resolving issues with latency and processing efficiency, is another possible direction for further study.