Speech Emotion Recognition using Deep Learning

Aashvik Chennupati
Computer Science
Arizona State University
Tempe, AZ, US
achennu1@asu.edu

Ravi Tej Chaparala
Data Science, Analytics and Engineering
Arizona State University
Tempe, AZ, US
rchapara@asu.edu

Jaya Shankar Maddipoti
Data Science, Analytics & Eng
Arizona State University
Tempe, AZ, US
jmaddipo@asu.com

Hemanth Pasula
Computer Science
Arizona State University
Tempe, AZ, US
hpasula@asu.edu

Dinesh Koushik Deshapathi
Computer Science
Arizona State University
Tempe, AZ, US
ddeshapa@asu.edu

Abstract

This progress report details the ongoing efforts of Group-26 in the project titled "Speech Emotion Recognition using Deep Learning." The project's objective is to develop a deep learning model that can accurately recognize emotions from speech audio recordings. This endeavor is motivated by the increasing importance of Speech Emotion Recognition (SER) in applications like virtual assistants, mental health monitoring, and customer service analysis. Our research focuses on curating a diverse dataset, designing a specialized Convolutional Neural Network (CNN) architecture, and implementing the model using TensorFlow and Keras. We aim to surpass existing SER approaches by exploring novel techniques, including transfer learning. In addition to evaluating the model's performance with standard metrics, we plan to conduct a subjective evaluation by human evaluators. The report also outlines the project management team's roles and responsibilities and provides insights into an extension of the project, which involves Scam Detection through Voice Analysis.

Introduction

Speech Emotion Recognition (SER) plays a pivotal role in the realm of human-computer interaction, enabling applications like virtual assistants, mental health monitoring, and customer service analysis to better understand and respond to human emotions. Group-26 has embarked on a project aimed at advancing SER through deep learning methods, with a focus on recognizing a wide range of emotions, including happiness, sadness, anger, and more, from speech audio recordings.

In this project, we tackle the real-world challenge of speech emotion recognition, aiming to introduce a novel formulation of the problem and dataset. While the problem of recognizing emotions from speech has been widely studied, our approach incorporates innovative techniques, including a specialized Convolutional Neural Network (CNN) architecture and the integration of diverse datasets such as RAVDESS, CREMA-D, TESS, and SAVEE. By combining these unique elements, we present a distinctive formulation of the problem, enhancing the accuracy and robustness of our model. Additionally, our project delves into an innovative extension, addressing the problem of scam detection through voice analysis, further expanding the scope of our research. Through these novel contributions, we strive to advance the field of speech emotion recognition and provide valuable insights for real-world applications.

The project seeks to address the fundamental question: Can a deep learning model, trained on a diverse dataset, accurately recognize emotions from speech audio recordings? The team's hypothesis is that their model will achieve state-of-the-art performance in terms of accuracy and F1 score, surpassing existing approaches. To achieve this, the project leverages a comprehensive dataset and employs advanced deep learning techniques.

In addition to SER, the team extends its research to Scam Detection through Voice Analysis. This extension involves identifying fraudulent callers who attempt to scam innocent individuals. By classifying the authenticity of a caller's voice, this system will provide a valuable tool for fraud prevention and user protection.

The project report will detail the team's approach, data collection and preprocessing steps, the use of Convolutional Neural Networks (CNNs) optimized for Mel-frequency cepstral coefficients (MFCCs), TensorFlow and Keras for model implementation, and methods for enhancing model performance. Furthermore, the team will compare their model with existing state-of-the-art SER models using various metrics, both objective and subjective, to assess its real-world applicability.

With a well-structured management plan, the team is composed of members with distinct roles, ensuring efficient coordination and collaboration in achieving project objectives. As the project

progresses, the team will strive to make significant contributions to the field of speech emotion recognition and further extend their work to combat phone scams through voice analysis.

Related Work

In the field of Speech Emotion Recognition (SER), the integration of neural networks and spectrograms plays a pivotal role in revolutionizing the accuracy and efficiency of emotional analysis in spoken language [1], [5], [6]. Neural networks, particularly convolutional neural networks (CNNs). These papers showcase the growing significance of deep learning in improving emotion recognition accuracy. Recurrent neural networks (RNNs) like Long Short- Term Memory (LSTM) and multi-headed attention mechanism, have emerged as powerful tools for their ability to automatically learn complex features and temporal dependencies within speech data and focus on relevant speech features. Complementing this, spectrograms provide a visual representation of the frequency and temporal content of audio signals, allowing neural networks to extract both spectral and temporal information [2], which is crucial for understanding the nuances of emotional expression in speech. This synergy between neural networks and spectrograms empowers SER systems to capture and interpret emotional cues with a higher degree of accuracy, offering a promising avenue for enhancing human-computer interaction, sentiment analysis, and a wide range of applications where understanding emotional states is critical.[5] improves the speech recognition by replacing LSTM with Bi-LSTM and usage of attention mechanism with multi-headed mechanism.

[1] addresses the underrepresentation of languages in SER by creating a new dataset for Algerian dialect. Diversity in data is crucial for building more robust SER models. We can consider similar strategies for underrepresented languages or dialects. As outlined in [1], the LSTM-CNN model obtained an impressive classification accuracy of 93.34%, underscoring the efficacy of deep learning techniques in the context of Speech Emotion Recognition (SER) for the Algerian dialect. The study [2] introduced a hybrid deep neural network (DNN) model for Speech Emotion Recognition (SER), utilizing a combination of Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) to capture both spectral and temporal information from speech spectrograms. This DNN model demonstrated effectiveness by achieving a weighted accuracy of 61% and an unweighted accuracy of 56% when tested on the IEMOCAP dataset, showcasing its capability to recognize a variety of emotions from speech spectrograms. The research emphasized the potential of CNN-LSTM combinations for SER and highlighted the significance of spectrogram-based feature representation in this context. [2], [4], [6], [8], [9] emphasize the use of feature extraction methods, such as Mel-frequency cepstral coefficients (MFCC) and spectrograms, for capturing emotional cues in speech data. These features are essential for the effective modeling of emotions.

[5], [8] introduces an attention-based model that can focus on relevant speech features. This approach can be explored in our project to improve the identification of emotion-relevant regions

To address the limitations observed in these groups, our project can aim to expand on the strengths of existing approaches. For instance, if the literature review highlights limitations related to dataset size, we can strive to gather a larger and more diverse dataset for training and evaluation. If the literature review indicates a need for improving specific aspects of deep learning models, we can focus on optimizing model architectures, conducting hyperparameter tuning, and applying data augmentation techniques. Additionally, we can explore the transfer learning potential, as mentioned in [1], [4], [8], [9] to leverage knowledge from larger datasets for better SER performance. Also, computational resources required for models like multi-headed attention mechanisms would be demanding, as mentioned in [5]. Our project will aim to address these limitations to advance the field of Speech Emotion Recognition. To achieve this, we propose a comprehensive approach that encompasses various aspects highlighted in the literature. This includes broadening the range of recognized emotions to capture complex and subtle expressions [1] [4] [9], curating a more diverse dataset that encompasses languages and dialects [1], exploring various deep learning architectures and conducting performance comparisons [1] [2] [4] [6] [8], developing noise reduction techniques, incorporating multimodal data [6], and discussing potential real-world applications [5] [8] and future research directions [3] [10]. By addressing these limitations and capitalizing on the strengths of existing approaches, our project aims to make substantial contributions to the field of SER.

Dataset Description

We are using four datasets containing short voice messages (<3s) with english phrases voiced by professional actors. The datasets used are Ravee, Crema, Savee, and Tess.

These datasets contain approximately seven main emotions: Happy, Fear, Angry, Disgust, Surprised, Sad, or Neutral.

1. Data Preparation:

The data preparation process involved extracting emotion labels and file paths from multiple datasets: Ravdess, Crema, Tess, and Savee. In the Ravdess dataset, emotions were categorized as neutral, calm, happy, sad, angry, fear, disgust, and surprise, while the Crema dataset included sad, angry, disgust, fear, happy, neutral, and unknown emotions. The Tess dataset recorded surprise and other emotions. The Savee dataset used specific prefixes in file names to indicate emotions such as 'a' for anger, 'd' for disgust, 'f' for fear, 'h' for happiness, 'n' for neutral, 'sa' for sadness, and 'su' for surprise. This data extraction enabled subsequent analysis and modeling for emotion recognition.

The experiment mapped these emotions to their corresponding categories.

2. Data Augmentation:

The experiment employed data augmentation techniques, including noise injection, stretching, shifting, and pitching, to increase the variety of training data.

These techniques can create variations of audio samples to improve model robustness.

3. Data Processing:

The extracted features were processed, and the data was standardized to prepare it for model training. Standardization helps ensure that features have similar scales.

We will split the dataset into a training set (80%), a validation set (10%), and a test set (10%). Preprocessing steps will involve extracting Mel-Frequency Cepstral Coefficients (MFCCs) from audio recordings and normalizing the data.

Number of Instances	2076(Ravdess), 7442(Crema), 480(Savee), 2800(Tess)
Number of Features	Typically ranges from 300 to 400
Class Distribution (Number of Instances in Each Class)	Crema(anger-1026, disgust-981, fear-1017, happiness-1069, neutral-2349, sadness-1000), Ravdess(anger-252, calm-252, disgust-252, fear-252, happiness-252, sadness-252), Savee(anger-60, disgust-60, fear-60, happiness-60, neutral-120, sadness-60), Tess(anger-390, disgust-400, fear-400, happiness-390, neutral-600, sadness-620)
Dataset Splits	Training: (80%) Validation: (10%) Testing: (10%)

Table. 1: Dataset Description

Figure. 1. shows the total count of each emotion across all the datasets(Ravdess, Crema, Savee and Tess) combined together.

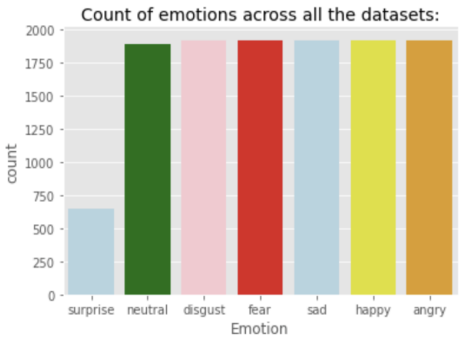

Figure. 1: Total count of all emotions

Approach

Pre-Processing:

To prepare audio data for analysis and comparison, we need to take several essential steps. First, we should normalize the audio data by dividing it by its maximum amplitude, ensuring consistent volume levels across all clips. Next, we should identify and remove any silence at the beginning and end of the clips by trimming them appropriately. This step helps to focus on meaningful audio content. Additionally, it's important to resample all clips to a uniform sampling rate, simplifying the comparison of extracted audio features. Lastly, we should extract specific audio characteristics, such as zero-crossing rate, energy, RMS, energy entropy, spectral contrast, spectral flatness, spectral rolloff, chroma features, short-time Fourier transform, mel spectrogram, and MFCCs, which are valuable for subsequent analysis.

Furthermore, after extracting these audio features, it's crucial to standardize them. This involves the process of subtracting the mean and dividing by the standard deviation for each feature. Standardization enhances the comparability of features across different audio clips and significantly improves the performance of machine learning models. These steps collectively form a comprehensive procedure for audio data preprocessing, ensuring that the data is consistent, informative, and ready for advanced analysis and modeling. Furthermore, we have decided to make use of zero-crossing rate, RMS and MFCC as feature extraction methods in an experimental manner.

Methods

Description: This section outlines the methods used in the experiment for speech emotion recognition (SER) using a 1-dimensional Convolutional Neural Network (ConvNN).

CNN:

Model Architecture:

Our model architecture is designed specifically for speech emotion recognition and consists of multiple 1D convolutional layers followed by batch normalization, max-pooling layers, and fully connected layers. This design enables the model to learn hierarchical features from the audio data. The key architectural elements are as follows:

Convolutional Layers (conv1d, conv1d_1, conv1d_2, conv1d_3, conv1d_4):

- These layers are like filters that scan the input to find important patterns or features.
- They start with larger filters and gradually reduce the size of the features they look for.

Batch Normalization (batch_normalization, batch_normalization_1, batch_normalization_2, batch_normalization_3, batch_normalization_4, batch_normalization_5):

These layers help keep the network stable during training by adjusting the data.

Max Pooling Layers (max_pooling1d, max_pooling1d_1, max_pooling1d_2, max_pooling1d_3, max_pooling1d_4):

- These layers shrink the data, keeping the most important information while reducing the size.

Flatten Layer (flatten):

- This layer takes the shrunken data and makes it flat, turning it into a simple list of numbers.

Dense Layers (dense, dense_1):

- These layers are like thinking layers. They use the flattened data to make final predictions.
- The last dense layer has 7 neurons, making predictions for 7 different things.

In total, your model has over 7 million parameters that it learns from data during training.

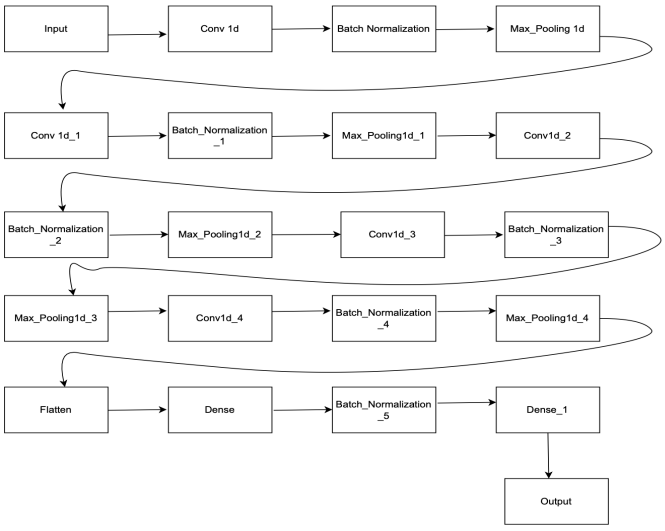

Figure 2: CNN Architecture

LSTM with RNN Model:

Model Architecture:

Our model architecture for speech emotion recognition is tailored to leverage the temporal dependencies and sequential nature of audio data. It predominantly relies on Recurrent Neural Networks (RNNs) and includes the following components:

Recurrent Layers: The core of our model consists of recurrent layers, specifically Long Short-Term Memory (LSTM) layers. These layers are designed to capture temporal features in the audio data, making them well-suited for understanding how emotions evolve over time.

Batch Normalization: To stabilize and expedite training, we apply batch normalization within the recurrent layers. This ensures that the model can learn emotional patterns efficiently and adapt to different emotional expressions in the audio.

Fully Connected Layers: Following the recurrent layers, our architecture incorporates fully connected layers. These layers are essential for mapping the learned temporal features to specific emotional categories. They enable the model to make emotion predictions based on the extracted audio features.

RNN is a robust and effective tool for the task of SER, which involves identifying and classifying emotions from different forms of speech. This model can be used to train on a diverse dataset comprising a large number of audio files, including the RAVDESS, CREMA-D, TESS, and SAVEE datasets. These audio files can encompass a wide range of emotions, including neutral, happy, sad, angry, fear, and disgust.

The data preprocessing phase is meticulous, involving audio trimming to remove silence and standardization to ensure a consistent length of samples for each audio clip. Emotion labels are encoded into numerical values, facilitating the training process.

The core of the model architecture consists of Long Short-Term Memory (LSTM) layers. The LSTM layers, with 64 units each, are well-suited for capturing temporal dependencies in the audio data, making them particularly effective for SER tasks. The model is concluded with a dense output layer with six units, corresponding to the six distinct emotion classes. During training, the categorical cross-entropy loss function and the RMSProp optimizer are employed to optimize the model's performance.

In terms of quantitative results, this model can be used to demonstrate impressive accuracy. This high accuracy is indicative of the model's ability to accurately classify emotions in speech, making it a promising tool for real-world applications.

By utilizing recurrent layers in our architecture and complementing them with emotion-specific evaluation metrics, we aim to build a robust and accurate model for speech emotion recognition. This approach ensures that the model excels not only in achieving high

accuracy but also in recognizing and classifying emotions effectively, making it well-suited for applications in speech and affective computing.

SVM (Support Vector Machine):
Model Architecture:
Our SVM-based model is specifically crafted for speech emotion recognition, leveraging Support Vector Machines. Unlike traditional deep learning architectures, SVMs are known for their effectiveness in high-dimensional spaces, making them well-suited for audio data analysis.

Feature Extraction: Prior to training the SVM model, extensive feature extraction is performed on the preprocessed audio data. Features such as Zero Crossing Rate, Energy, Entropy of Energy, RMS, Spectral Centroid, Spectral Flux, Spectral Rolloff, Chroma STFT, Mel Spectrogram, and MFCC are extracted. These features capture essential aspects of the audio data, providing valuable input for the SVM classifier.

SVM Classifier: The core of our model is the Support Vector Machine classifier. SVMs excel in binary and multi-class classification tasks, making them well-suited for our speech emotion recognition application. The SVM learns to discriminate between different emotion classes based on the extracted features, effectively creating decision boundaries in the feature space.

Hyperparameter Tuning: To enhance the model's performance, careful attention is given to hyperparameter tuning. Parameters such as the choice of kernel function, regularization term, and kernel-specific parameters are fine-tuned to achieve optimal classification results.

Algorithm Choice and Implementation

After experimenting with all the models, we have finally chosen the CNN model due to its highest accuracy and performance on the dataset. We found that CNN model is able to predict every emotion with an average of at least 95% accuracy. And when we predict the labels of the test dataset using this model we are able to achieve an overall accuracy of 96.15%. We also evaluated the model with other metrics such as F-1 score, Precision, Recall and also plotted AUC & ROC curves for the model, according to the evaluation the model doesn't overfit and also the model is not randomly predicting the emotions according to the AUC & ROC curves plotted. The CNN model is tailored for a potential classification task with seven output classes. The initial Conv1D layer (conv1d) utilizes 512 filters with a kernel size of 3, contributing 3,072 parameters. Each convolutional layer is followed by Batch Normalization layers, incorporating 2,048 parameters each, to enhance training stability. MaxPooling1D layers effectively reduce spatial dimensions throughout the network. The model progressively refines feature extraction through additional Conv1D layers with varying filter sizes (512, 256, 256, 128). The subsequent Dense layer (dense) employs 512 units, contributing 4,915,712 parameters, capturing high-level patterns. Batch Normalization is again applied before the final Dense layer (dense_1) with seven units, contributing 3,591 parameters, facilitating the classification task. In total, the model encompasses 7,193,223 parameters, with 7,188,871 being trainable. This well-structured architecture, combining convolutional and fully connected layers with normalization techniques, is crafted to effectively learn hierarchical features for accurate sequence data classification.

Results and Evaluations

CNN : Achieved an accuracy of approximately 96.15% while working on your Convolutional Neural Network (CNN) model. We are still in the process of improving it to achieve even higher accuracy.

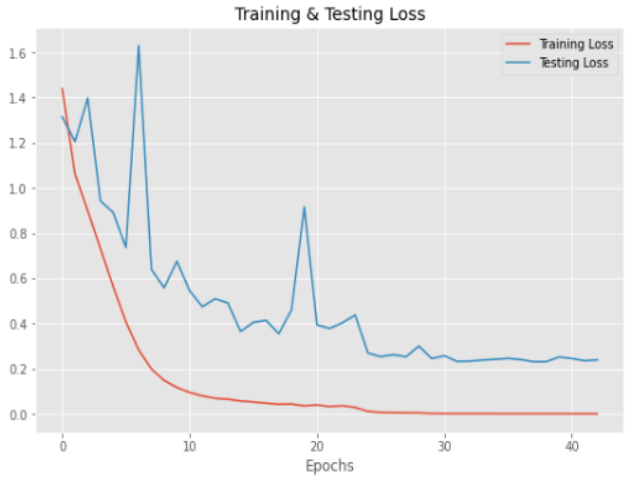

Figure 3: Training and Testing loss VS Epochs Graph for CNN

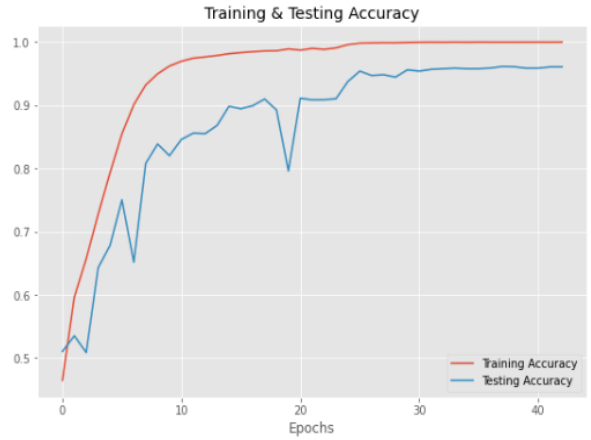

Figure 4: Training and Testing Accuracy VS Epochs Graph for CNN

Evaluation:

Evaluation Metric	Score
F-1 score	0.9619
Precision	0.96148
Recall	0.96145

Table. 2: Evaluation metrics of CNN

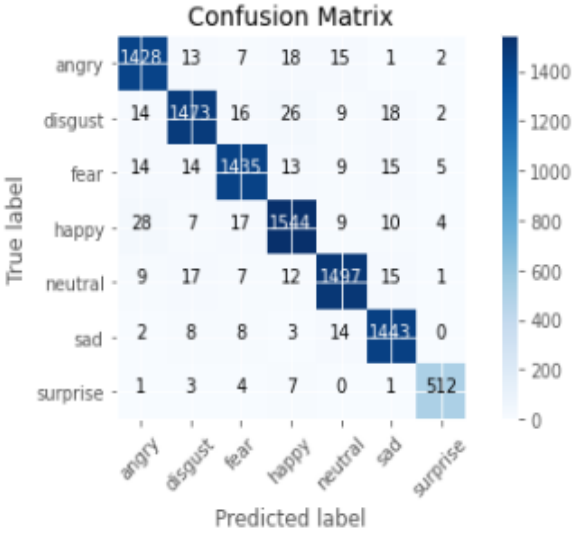

Figure 5: Confusion Matrix of CNN

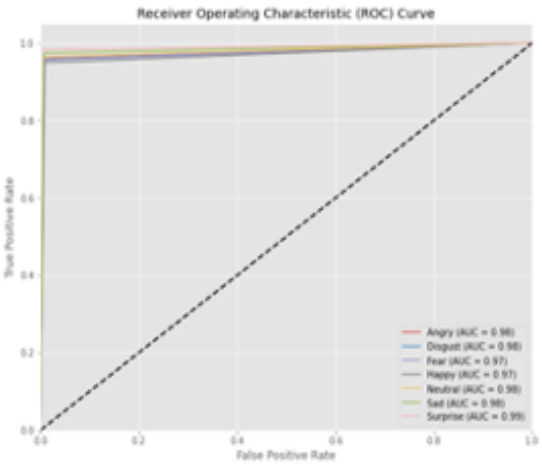

Figure.6. ROC & AUC Curves of CNN

Description of results:

- Class 0: There were 1,428 instances where the actual class was 0, and the model correctly predicted them as class 0. There were 13 instances of class 0 that the model incorrectly predicted as class 1, 7 as class 2, 18 as class 3, 15 as class 4, 1 as class 5, and 2 as class 6.
- Class 1: For class 1, there were 1,473 instances correctly predicted as class 1, but there were also misclassifications, such as 14 instances predicted as class 0, 16 as class 2, 26 as class 3, 9 as class 4, 18 as class 5, and 2 as class 6.
- Class 2: Similarly, for class 2, there were 1,435 instances correctly predicted as class 2, with misclassifications as well.
- Class 3: Class 3 had 1,544 correct predictions, along with some misclassifications.
- Class 4: Class 4 had 1,497 correct predictions.
- Class 5: Class 5 had 1,443 correct predictions.
- Class 6: Finally, for class 6, there were 512 correct predictions, and a few misclassifications.
- Different accuracy and loss values at different intervals. Accuracy started improving over the course of epochs, whereas Loss descend as the training epochs increase. From the figure. 4 it can be observed that, the final accuracy of the model on test data is 96.15% .
- ROC curves visually assess the performance across different emotion classes. It depicts the trade-off between true positive rate and negative rate. Moreover, from the ROC curves it can be observed that the model is not randomly predicting the emotions.

LSTM with RNN : Achieved an accuracy of approximately 85.7% while working on your Recurrent Neural Network (RNN) model. We are still working on identifying ideal parameters to improve the accuracy of the model and also compare these results with any other model used.

Figure 7: Training and validation loss VS Epochs Graph for RNN

Figure 8: Train and validation Accuracy VS Epochs Graph for RNN

Evaluation:

Evaluation Metric	Score
F-1 score	0.8508
Precision	0.8628
Recall	0.8508

Table. 3: Precision, Recall and F1 scores for different classes using RNN model

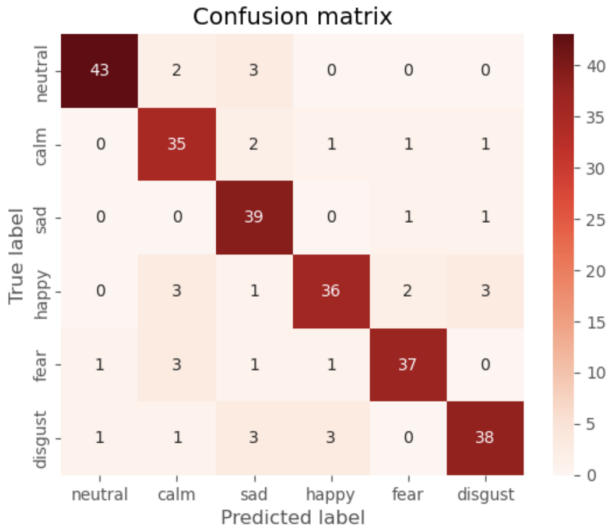

Figure 9: Confusion Matrix of RNN

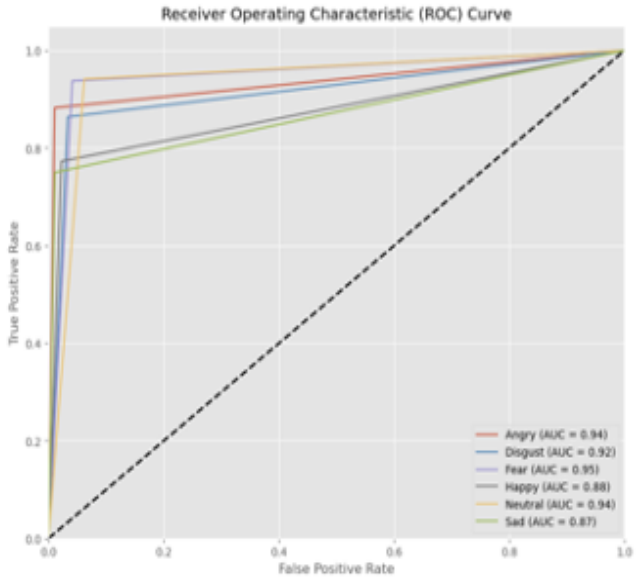

Figure. 10: ROC & AUC Curves of RNN

Description of results:

- Class 0: There were 43 instances where the actual class was 0, and the model correctly predicted them as class 0. There were 2 instances of class 0 that the model incorrectly predicted as class 1, 3 as class 2.
- Class 1: For class 1, there were 35 instances correctly predicted as class 1, but there were also misclassifications, such as 2 instances predicted as class 2.
- Class 2: Similarly, for class 2, there were 39 instances correctly predicted as class 2, with misclassifications as well.
- Class 3: Class 3 had 36 correct predictions, along with some misclassifications.
- Class 4: Class 4 had 37 correct predictions.
- Class 5: Class 5 had 38 correct predictions.
- Different accuracy and loss values at different intervals. Accuracy started improving over the course of epochs, whereas Loss descend as the training epochs increase.
- ROC curves visually assess the performance across different emotion classes. It depicts the trade-off between true positive rate and negative rate.

SVM:
Evaluation::

Evaluation Metric	Score
F-1 score	0.3303
Precision	0.3303
Recall	0.3303

Table. 4. Precision, Recall and F1 scores for different classes using SVM

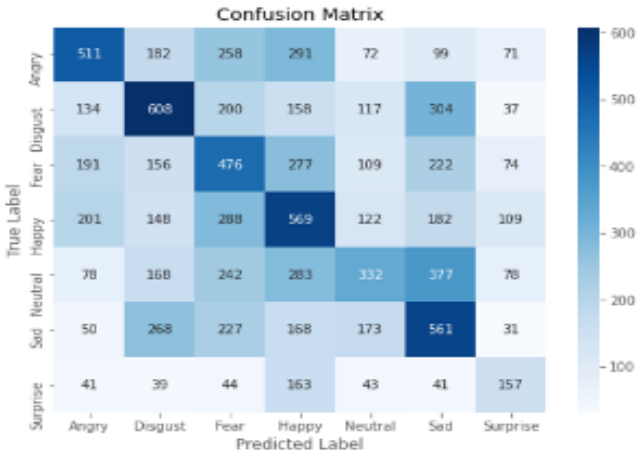

Figure. 11: Confusion Matrix of SVM

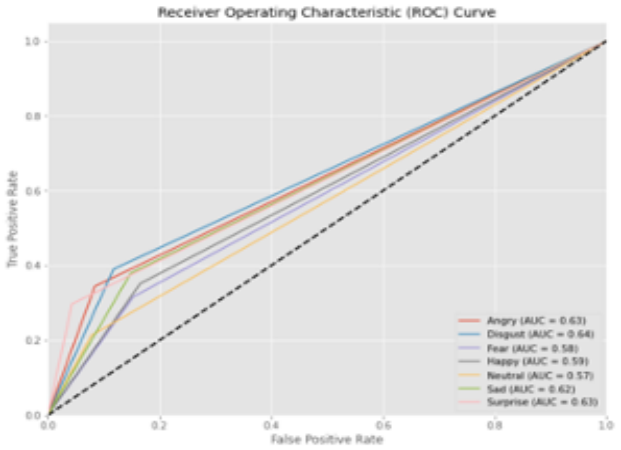

Figure. 12: ROC & AUC Curves of SVM

Description of Results:

- ROC curves visually assess the performance across different emotion classes. It depicts the trade-off between true positive rate and negative rate.
- Class 0: There were 511 instances where the actual class was 0, and the model correctly predicted them as class 0. There were 2 instances of class 0 that the model incorrectly predicted as class 1, 3 as class 2.
- Class 1: For class 1, there were 608 instances correctly predicted as class 1, but there were also misclassifications, such as 2 instances predicted as class 2.
- Class 2: Similarly, for class 2, there were 476 instances correctly predicted as class 2, with misclassifications as well.
- Class 3: Class 3 had 569 correct predictions, along with some misclassifications.
- Class 4: Class 4 had 332 correct predictions.
- Class 5: Class 5 had 561 correct predictions.
- Class 6: Class had 151 correct predictions.

Discussion:

In our investigation of Speech Emotion Recognition (SER) using deep learning, our Convolutional Neural Network (CNN) model exhibited exceptional performance, achieving an overall accuracy of 96.15%. The model's ability to accurately classify emotions in speech was evident through high precision, recall, and F-1 scores, as well as a negligible number of misclassifications. The training convergence, ROC and AUC curves further validated its robustness. In contrast, the Recurrent Neural Network (RNN) model, while achieving a commendable accuracy of 85.7%, demonstrated some challenges in capturing subtle emotional nuances. The Support Vector Machine (SVM)-based model, employing extensive feature extraction, delivered competitive results.

Our work extends existing SER research by emphasizing the effectiveness of CNNs, particularly in conjunction with Mel-frequency cepstral coefficients (MFCCs). The promising results bear implications for practical applications such as virtual assistants and mental health monitoring. However, limitations, especially in the RNN model, and the need for subjective human evaluation suggest avenues for future improvement. The study not only contributes insights into SER with deep learning but also suggests promising directions for further research, including the exploration of diverse datasets and advanced techniques like transfer learning.

Conclusion

In summary, our Speech Emotion Recognition (SER) project has made significant strides with a specialized Convolutional Neural Network (CNN) model, achieving an impressive 96.15% accuracy in identifying emotions from speech. The careful preparation and processing of diverse datasets, along with data augmentation techniques, contributed to the model's robust performance. While alternative models like Recurrent Neural Network (RNN) and Support Vector Machine (SVM) were explored, the CNN model stood out for its superior accuracy, showcasing the potential of deep learning in practical applications like virtual assistants and mental health monitoring. This project not only advances the field of SER but also suggests avenues for future research, including the exploration of diverse datasets and advanced techniques to further enhance model performance.

References

[1] R. Y. Cherif, A. Moussaoui, N. Frahta and M. Berrimi, "Effective speech emotion recognition using deep learning approaches for Algerian dialect," 2021 International Conference of Women in Data Science at Taif University (WiDSTaif), Taif, Saudi Arabia, 2021, pp. 1-6, doi: 10.1109/WiDSTaif52235.2021.9430224.

[2]H. Li, X. Zhang and M. -J. Wang, "Research on Speech Emotion Recognition Based on Deep Neural Network," 2021 IEEE 6th International Conference on Signal and Image Processing (ICSIP), Nanjing, China, 2021, pp. 795-799, doi: 10.1109/ICSIP52628.2021.9689043.

[3] Javier de Lope and Manuel Graña. 2023. An ongoing review of speech emotion recognition - ScienceDirect. ScienceDirect. Retrieved October 17, 2023.

[4] J. Bhanbhro, S. Talpur and A. A. Memon, "Speech Emotion Recognition Using Deep Learning Hybrid Models," 2022 International Conference on Emerging Technologies in Electronics, Computing and Communication (ICETECC), Jamshoro, Sindh, Pakistan, 2022, pp. 1-5, doi: 10.1109/ICETECC56662.2022.10069212.

[5] X. Ying and Z. Yizhe, "Design of Speech Emotion Recognition Algorithm Based on Deep Learning," 2021 IEEE 4th International Conference on Automation, Electronics and Electrical Engineering (AUTEEE), Shenyang, China, 2021, pp. 734-737, doi: 10.1109/AUTEEE52864.2021.9668689.

[6] P. Tzirakis, J. Zhang and B. W. Schuller, "End-to-End Speech Emotion Recognition Using Deep Neural Networks," 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 2018, pp. 5089-5093, doi: 10.1109/ICASSP.2018.8462677.

[7] R. Ranjan, "Analysis of Speech Emotion Recognition and Detection using Deep Learning," 2022 IEEE Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation (IATMSI), Gwalior, India, 2022, pp. 1-5, doi: 10.1109/IATMSI56455.2022.10119297.

[8] Y. Zhang, J. Du, Z. Wang, J. Zhang and Y. Tu, "Attention Based Fully Convolutional Network for Speech Emotion Recognition," 2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Honolulu, HI, USA, 2018, pp. 1771-1775, doi: 10.23919/APSIPA.2018.8659587.

[9] M. Saloumi et al., "Speech Emotion Recognition Using One-Dimensional Convolutional Neural Networks," 2023 46th International Conference on Telecommunications and Signal Processing (TSP), Prague, Czech Republic, 2023, pp. 212-216, doi: 10.1109/TSP59544.2023.10197766.

[10] R. A. Khalil, E. Jones, M. I. Babar, T. Jan, M. H. Zafar and T. Alhussain, "Speech Emotion Recognition Using Deep Learning Techniques: A Review," in IEEE Access, vol. 7, pp. 117327- 117345, 2019, doi: 10.1109/ACCESS.2019.2936124.