Enhancing Audio-Based Emotion Recognition through Feature Fusion and Model Refinement

Ruilin Wu
School of Electrical and Computer Engineering
The University of Sydney
Sydney, Australia
SID:520035056
ruilin.wu@sydney.edu.au

***Abstract*—The recognition and understanding of human emotions through spoken language is a pivotal aspect of enhancing human-computer interaction. Emotions, being a reflection of a human's mental state, play a crucial role in conveying one's outlook and feelings. While humans naturally discern emotions in speech, machines require sophisticated techniques to achieve this. This paper delves into the realm of Speech Emotion Recognition (SER), a significant research field that aims to predict human emotions from their speech patterns. Using the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) dataset and Toronto emotional speech set data (TESS), we extracted key audio features such as Mel Frequency Cepstral Coefficients (MFCC), Chroma, Spectral Contrast and others to train various machine learning classifiers including GRU, LSTM, Random Forests, and CNNs. Despite the advancements, challenges persist, such as dealing with variations and noise in data, and the subjective nature of emotions. However, the results underscore the potential of deep learning and feature extraction in advancing the field of emotion recognition. For further research and replicability, the source code and methodologies are made available on our GitHub repository: https://github.com/ruilin-wu/Emotion_Recognition_with_Enhancing_Models.**

***Index Terms*—Speech Emotion Recognition, deep learning, classification, Mel Frequency Cepstral Coefficients, Convolutional Neural Networks.**

I. Introduction

In the realm of human-computer interaction, understanding human emotions through spoken language has emerged as a pivotal area of research. The ability to recognize and interpret emotions from speech can significantly enhance the quality of interaction between humans and machines, making the latter more empathetic and responsive to the user's needs. This capability is not just limited to humans; even animals, such as dogs and elephants, have shown the ability to discern human emotions through vocal cues [1].

The process of predicting human emotions through speech, termed as "Speech Emotion Recognition (SER)", has garnered considerable attention in recent years. Despite the inherent challenges, such as the subjectivity of emotions and the complexities in annotating audio data, advancements in this field have been promising [1]. Various features of speech, including tone, pitch, and expression, serve as indicators of the underlying emotion. Among these, certain features like the Mel Frequency Cepstral Coefficients (MFCC) have been extensively used for emotion recognition [3].

Several datasets, like RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song), have been instrumental in facilitating research in this domain. The RAVDESS dataset, for instance, comprises audio files from various individuals, each representing distinct emotions [1]. Utilizing such datasets, researchers have employed a range of machine learning algorithms, from traditional ones like Support Vector Machines (SVM) to more advanced deep learning models like Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks, to classify emotions [2][3][4].

However, the journey of SER is not without its challenges. The variability in speech patterns, influenced by factors such as cultural background, individual personality, and context, can pose difficulties in accurate emotion detection. Moreover, while some algorithms have shown promise in lab-controlled environments, their performance in real-world scenarios remains to be thoroughly evaluated.

In light of these challenges and the potential benefits of SER, this paper aims to delve deeper into the intricacies of emotion recognition from speech. Drawing insights from the seminal work of M. G. de Pinto et al., 2020, and complementing it with findings from other notable studies in the field, this paper endeavors to present a comprehensive overview of this field.

II. Literature Review

In recent years, the field of Speech Emotion Recognition (SER) has garnered significant attention, primarily due to its potential in enhancing human-computer interaction. SER aims to predict human emotions through their speech, making it a challenging yet crucial domain in the realm of audio signal processing and machine learning.

One of the pioneering works in this area was presented by de Pinto et al. [1]. Their research emphasized the importance of Mel-Frequency Cepstral Coefficients (MFCC) in understanding emotions from spoken language. Using deep neural networks, they achieved promising results, highlighting the potential of MFCC as a robust feature for SER. Their work laid the foundation for many subsequent studies, emphasizing the importance of feature extraction and the choice of appropriate machine learning models.

Hussain et al. [2] further explored the domain of audio-visual emotion recognition. They emphasized that emotions play a pivotal role in deciphering a human's mental state. Their research combined auditory and visual cues, introducing novel speech features like the Wavegram. This feature, extracted using a one-dimensional Convolutional Neural Network (CNN) from time-domain waveforms, showcased superior performance compared to traditional methods.

Another noteworthy contribution came from Andayani et al. [3], who introduced a hybrid model combining the strengths of Long-Short Term Memory (LSTM) and Transformer architectures. Their model, trained on the RAVDESS dataset, achieved impressive accuracies, underscoring the effectiveness of combining different deep learning architectures for SER.

Deshmukh et al. [4] and Mittal et al. [5] both delved into the machine learning aspect of SER. While Deshmukh et al. explored various machine learning algorithms like Support Vector Machines (SVM) and Random Forests, Mittal et al. focused on the feature extraction process, emphasizing the significance of tone, pitch, and energy in speech signals. Both studies utilized multiple datasets, including SAVEE, RAVDESS, and TESS, to train and test their models.

Anusha et al. [6] further expanded on the machine learning techniques for SER. Their research highlighted the potential of classifiers like Recurrent Neural Network (RNN), K-Nearest Neighbors (k-NN), and Multi-Layer Perceptron (MLP). Their experiments revealed that the MLP classifier, when trained on features like MFCC and Chroma, achieved the highest accuracy among all tested classifiers.

In conclusion, the field of Speech Emotion Recognition has seen rapid advancements in recent years. From feature extraction techniques like MFCC and Wavegram to machine learning models like CNN, LSTM, and MLP, researchers have explored various avenues to enhance the accuracy and reliability of SER systems. As technology continues to evolve, it is anticipated that SER will play an even more integral role in human-computer interactions, making our devices more empathetic and responsive to our emotional states.

III. Dataset Summary

In this study, the dataset was compiled from 5252 samples, sourced from two primary databases [1]:

From the RAVDESS database, the dataset incorporated 1440 speech files and 1012 song files [7]. This collection features recordings by 24 professional actors, comprising 12 females and 12 males. They vocalized two statements that matched in terms of lexicon, delivered in a neutral North American accent. The speech recordings encompassed various emotions such as calm, happy, sad, angry, fearful, surprise, and disgust. Meanwhile, the song recordings covered emotions like calm, happy, sad, angry, and fearful. Each of these files underwent a rating process, evaluated 10 times based on emotional validity, intensity, and authenticity. The ratings were given by 247 individuals, typical of untrained adult participants from North America. An additional group of 72 participants contributed data for test-retest reliability. The validation data, which is open-access, revealed high scores in emotional validity, interrater reliability, and test-retest intrarater reliability. Table 1 shows the total number of audio files for various emotions in the RAVDESS dataset.

TABLE I: Count of Speech and Song Files in RAVDESS Dataset

Emotion	Sample Count		Summed Count
	Speech	Song	
Neutral	96	92	188
Happy	192	184	376
Angry	192	184	376
Sad	192	184	376
Calm	192	184	376
Fearful	192	184	376
Surprised	192	0	192
Disgust	192	0	192
Total Count	1440	1012	2452

From the TESS database, the dataset integrated 2800 files [1]. Two actresses, aged 26 and 64, vocalized 200 target words within the carrier phrase "Say the word ". These recordings captured seven distinct emotions: anger, disgust, fear, happiness, pleasant surprise, sadness, and neutral, resulting in a total of 2800 stimuli. Both actresses, hailing from Toronto, are native English speakers with university education and musical training backgrounds. Audiometric tests confirmed that both had normal hearing thresholds.

The objective of the model is to classify the recordings into the following emotion categories: 0 for neutral, 1 for calm, 2 for happy, 3 for sad, 4 for angry, 5 for fearful, 6 for disgust, and 7 for surprised [1]. It's important to note that the dataset is imbalanced. The TESS database lacks the calm category, leading to fewer samples for that emotion. This imbalance becomes evident when reviewing the classification report.

IV. The Proposed Method

A. Dataset Selection and Processing

In the RAVDESS dataset, there are two major subsets: speech and song. The primary challenge in this study is to decide whether to train the model on just one subset or on the combined collection of both subsets. Within the song subset, due to the need for actors to express emotions using a wider range of tones, there's a valid concern that the noise might adversely affect the training accuracy of the model. As a result, we will create datasets that include only speech, only song, and a combination of both. Three separate models will be trained and evaluated based on these datasets. For evaluation, we will employ the LSTM model, with MFCC as the chosen feature.

TABLE II: Classification Report for Song Subset

Emotion	Metrics			support
	precision	**recall**	**F1-score**	
Neutral	0.97	0.99	0.98	165
Calm	0.81	0.87	0.84	62
Happy	0.92	0.87	0.89	190
Sad	0.88	0.92	0.90	183
Angry	0.93	0.93	0.93	189
Fearful	0.91	0.88	0.90	226
Disgust	0.97	0.98	0.98	125
Surprised	0.97	0.96	0.97	118
accuracy			0.92	1258
macro avg	0.92	0.93	0.92	1258
weighted avg	0.92	0.92	0.92	1258

TABLE III: Classification Report for Speech Subset

Emotion	Metrics			support
	precision	**recall**	**F1-score**	
Neutral	0.83	0.85	0.84	164
Calm	0.58	0.55	0.56	69
Happy	0.76	0.72	0.74	201
Sad	0.87	0.74	0.80	200
Angry	0.84	0.77	0.81	195
Fearful	0.73	0.88	0.79	194
Disgust	0.80	0.81	0.80	183
Surprised	0.76	0.80	0.78	194
accuracy			0.78	1400
macro avg	0.77	0.77	0.77	1400
weighted avg	0.79	0.78	0.78	1400

TABLE IV: Classification Report for Speech and Song Compilation

Emotion	Metrics			Support
	Precision	**Recall**	**F1-Score**	
Neutral	0.89	0.85	0.87	191
Calm	0.73	0.70	0.71	123
Happy	0.84	0.77	0.80	256
Sad	0.80	0.76	0.78	270
Angry	0.80	0.81	0.81	275
Fearful	0.71	0.84	0.77	228
Disgust	0.79	0.76	0.78	181
Surprised	0.80	0.81	0.81	210
Accuracy			0.79	1734
Macro Avg	0.79	0.79	0.79	1734
Weighted Avg	0.80	0.79	0.79	1734

Tables 2, 3, and 4 show classification results for three types of data: songs, speeches, and a mix of both. For songs, the numbers are higher, meaning the model works well. For speeches, the numbers are a bit lower, showing it might be harder to tell emotions from spoken words than from music. The mixed data has numbers in the middle.

Figure 1 displays the validation accuracy for 150 training steps for these data types. The song data starts off well and stays good, meaning the model finds it easier to learn from. The speech data starts lower but gets better, showing the model learns more as it trains. The mixed data goes up and down, probably because it has both songs and speeches. This can mean the model keeps changing its approach to get better results. The high numbers in the song data show that either the data is clearer or the model is just better suited for it. All models get better with more training, but the song model is clearly the best.

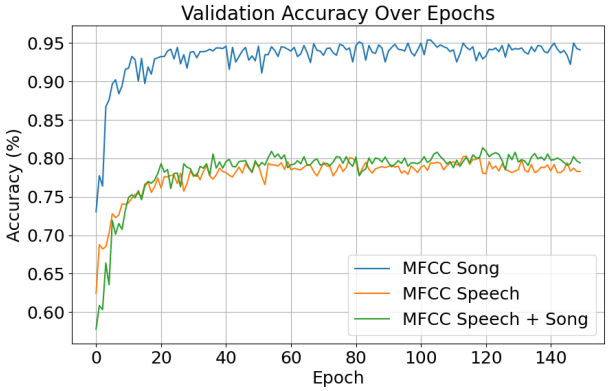

Fig. 1: Validation Accuracy Over Epochs for Different Dataset

B. Feature Extraction

In this section, this article will extract various relevant and meaningful features from the audio signal, namely MFCC, Chroma and Contrast, and introduce them into the model respectively to obtain the most meaningful features.

MFCCs (Mel-Frequency Cepstral Coefficients): MFCCs are derived from the cepstral representation of an audio clip. They reflect the short-term power spectrum of sound and are particularly effective in voice and speech recognition due to their ability to capture the phonetic content. The processing involves filtering the signal into Mel-frequency bands and subsequently using the Discrete Cosine Transform to obtain the coefficients.

Chroma: Chroma features capture the energy distribution across pitch classes and are useful in representing harmonic content. They relate to the twelve different pitch classes in western tonal music. By analyzing the intensity of each pitch class, these features can provide insight into the harmonic structure and progression within an audio segment.

Spectral Contrast: This feature measures the difference in amplitude between peaks and valleys in a sound spectrum. It offers information about the spectral shape of an audio signal, highlighting prominent frequency components against the sonic backdrop. Spectral contrast can be instrumental in distinguishing between harmonic and non-harmonic content in a sound segment.

Next, we will use the ”Speech and Song Compilation” database. We will extract features like MFCC, Chroma, and Contrast, as well as a combination of these three. Our goal is to analyze which features best represent human emotions. We will use the CNN model for this purpose [8].

TABLE V: Classification Report When Feature is Chroma

Emotion	Metrics			Support
	Precision	**Recall**	**F1-Score**	
Neutral	0.95	0.81	0.87	191
Calm	0.38	0.40	0.39	123
Happy	0.43	0.38	0.41	256
Sad	0.73	0.57	0.64	270
Angry	0.44	0.48	0.46	275
Fearful	0.62	0.49	0.55	228
Disgust	0.35	0.51	0.41	181
Surprised	0.41	0.51	0.46	210
Accuracy			0.52	1734
Macro Avg	0.54	0.52	0.52	1734
Weighted Avg	0.55	0.52	0.53	1734

TABLE VI: Classification Report When Feature is Spectral Contrast

Emotion	Metrics			Support
	Precision	**Recall**	**F1-Score**	
Neutral	0.87	0.61	0.72	191
Calm	0.22	0.11	0.15	123
Happy	0.34	0.43	0.38	256
Sad	0.64	0.39	0.49	270
Angry	0.50	0.59	0.54	275
Fearful	0.47	0.59	0.52	228
Disgust	0.38	0.60	0.47	181
Surprised	0.48	0.33	0.39	210
Accuracy			0.47	1734
Macro Avg	0.49	0.46	0.46	1734
Weighted Avg	0.50	0.47	0.47	1734

TABLE VII: Classification Report When Feature is MFCC

Emotion	Metrics			Support
	Precision	**Recall**	**F1-Score**	
Neutral	0.95	0.83	0.88	191
Calm	0.60	0.91	0.72	123
Happy	0.80	0.77	0.79	256
Sad	0.83	0.83	0.83	270
Angry	0.88	0.87	0.87	275
Fearful	0.80	0.80	0.80	228
Disgust	0.88	0.74	0.80	181
Surprised	0.84	0.85	0.85	210
Accuracy			0.82	1734
Macro Avg	0.82	0.83	0.82	1734
Weighted Avg	0.83	0.82	0.82	1734

From the table V-VIII, three distinct feature extraction methods have been used: Chroma, Spectral Contrast, and MFCC. A compound method, namely MFCC combined with Chroma and Contrast, has also been explored.

In the Chroma-based analysis (Table V), an overall accuracy of 0.52 is observed. Although a moderate performance, individual emotions like 'Neutral' show a higher F1-Score of 0.95. Contrastingly, in the Spectral Contrast method (Table VI), a slightly reduced accuracy of 0.47 emerges. Here, the F1-Scores across various emotions seem relatively homogenized, suggesting a consistent yet suboptimal performance.

However, the situation changes with the MFCC feature extraction. Table VII depicts an impressive accuracy of 0.82. The F1-Scores for emotions are predominantly in the high 0.7s and 0.8s, indicating a robust classification capability. The compounded method, MFCC + Chroma + Contrast, showcased in Table VII, confirms the superiority of MFCC as the combined methodology also reaches an accuracy of 0.82.

TABLE VIII: Classification Report When Feature is MFCC + Chroma + Contrast

Emotion	Metrics			Support
	Precision	**Recall**	**F1-Score**	
Neutral	0.96	0.81	0.88	191
Calm	0.69	0.88	0.77	123
Happy	0.88	0.75	0.81	256
Sad	0.86	0.75	0.80	270
Angry	0.91	0.85	0.88	275
Fearful	0.63	0.90	0.74	228
Disgust	0.86	0.79	0.82	181
Surprised	0.84	0.85	0.85	210
Accuracy			0.82	1734
Macro Avg	0.83	0.82	0.82	1734
Weighted Avg	0.84	0.82	0.82	1734

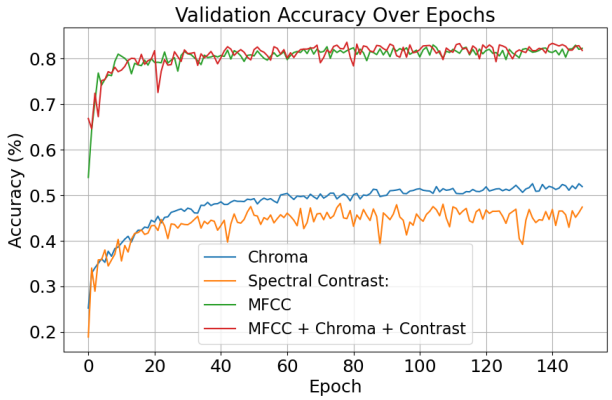

Fig. 2: Validation Accuracy Over Epochs for Different Features

Figure 2 complements tabular findings. While all methods show an increasing trend in accuracy across epochs, the convergence of the MFCC and MFCC + Chroma + Contrast methodologies towards higher accuracy is evident. Notably, the combined method exhibits promising stability in the later epochs.

To synthesize, the MFCC-based feature extraction, both in its standalone and combined form, clearly outperforms the other methodologies in the context of audio emotion classification.

C. Model Selection and Optimization

Based on previous research and discussions, we know that using only MFCC as features and using song subset training models can achieve better performance, so this paper proposes four machine learning models below to get the best performing model:

Basic CNN:

- **Conv1D Layer:** This is a 1-dimensional convolutional layer with 64 filters and a kernel size of 5. It is designed to learn spatial hierarchies from the input data.
- **Activation (ReLU):** The Rectified Linear Unit (ReLU) activation function introduces non-linearity to the model, allowing it to learn from the error and make adjustments, which is essential for learning complex patterns.
- **Dropout Layer:** This layer randomly sets a fraction (20% in this case) of the input units to 0 at each update during training, which helps to prevent overfitting.
- **Flatten Layer:** This layer flattens the multi-dimensional tensor into a 1-dimensional tensor to prepare the data for the Dense layer.
- **Dense Layer with Softmax Activation:** It's a fully connected layer that produces probabilities for each of the 8 categories.

Improved CNN:

- This model contains two sets of Conv1D layers, each followed by a Dropout layer. The first set has a kernel size of 3 and the second one has a kernel size of 4.
- **MaxPooling1D:** This layer down-samples the input representation, reducing its dimensionality and allowing for assumptions to be made about features contained in the sub-regions.
- The final layer is a Dense layer with softmax activation for classification into 8 categories.

1D CNN with GRU:

- This model integrates convolutional layers with a GRU (Gated Recurrent Unit) layer. GRUs are a type of recurrent neural network (RNN) that can remember past information and are used here to capture sequential dependencies.
- It starts with two Conv1D layers, similar to the improved CNN model.
- **GRU Layer:** GRUs are designed to learn patterns over time or sequences, which is particularly useful for time-series data or sequences.
- The final layer is again a Dense layer with softmax activation for classification.

LSTM Model:

- **LSTM Layers:** LSTM (Long Short-Term Memory) layers are a special kind of RNN that are capable of learning long-term dependencies. This model uses two LSTM layers.
- Similar to the previous models, Dropout layers are used to prevent overfitting.
- The final layer is a Dense layer with softmax activation for classification into 8 categories.

Next, this paper will evaluate these four models to determine the best-performing one.

TABLE IX: Classification Report for Basic CNN Model

Emotion	Metrics			**Support**
	Precision	**Recall**	**F1-Score**	
Neutral	0.99	1.00	1.00	165
Calm	0.92	0.77	0.84	62
Happy	0.88	0.94	0.91	190
Sad	0.89	0.96	0.92	183
Angry	0.89	0.97	0.93	189
Fearful	0.99	0.82	0.90	226
Disgust	1.00	0.99	1.00	125
Surprised	0.98	1.00	0.99	118
Accuracy			0.94	1258
Macro Avg	0.94	0.93	0.94	1258
Weighted Avg	0.94	0.94	0.94	1258

TABLE X: Classification Report for the Model

Emotion	**Metrics**			**Support**
	Precision	**Recall**	**F1-Score**	
Neutral	1.00	1.00	1.00	165
Calm	0.90	0.90	0.90	62
Happy	0.94	0.96	0.95	190
Sad	0.93	0.92	0.93	183
Angry	0.96	0.96	0.96	189
Fearful	0.94	0.92	0.93	226
Disgust	0.99	1.00	1.00	125
Surprised	0.98	0.99	0.99	118
Accuracy			0.96	1258
Macro Avg	0.96	0.96	0.96	1258
Weighted Avg	0.96	0.96	0.96	1258

TABLE XI: Classification Report for 1D CNN with GRU Model

Emotion	Metrics			**Support**
	Precision	**Recall**	**F1-Score**	
Neutral	0.99	0.98	0.98	165
Calm	0.96	0.79	0.87	62
Happy	0.92	0.93	0.92	190
Sad	0.92	0.85	0.88	183
Angry	0.92	0.97	0.94	189
Fearful	0.87	0.91	0.89	226
Disgust	0.98	0.99	0.98	125
Surprised	0.97	0.99	0.98	118
Accuracy			0.93	1258
Macro Avg	0.94	0.93	0.93	1258
Weighted Avg	0.93	0.93	0.93	1258

TABLE XII: Classification Report for LSTM Model

Emotion	Metrics			**Support**
	Precision	**Recall**	**F1-Score**	
Neutral	0.98	0.99	0.98	165
Calm	0.82	0.94	0.87	62
Happy	0.92	0.86	0.89	190
Sad	0.89	0.91	0.90	183
Angry	0.92	0.94	0.93	189
Fearful	0.88	0.85	0.87	226
Disgust	0.98	0.99	0.99	125
Surprised	0.97	0.97	0.97	118
Accuracy			0.92	1258
Macro Avg	0.92	0.93	0.92	1258
Weighted Avg	0.92	0.92	0.92	1258

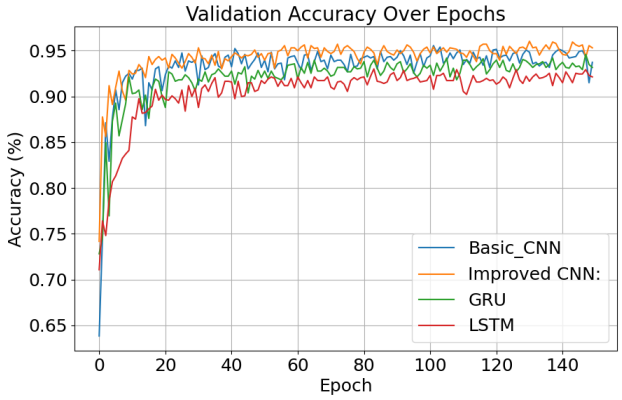

Fig. 3: Validation Accuracy Over Epochs for Different Models

From the table IX - XII, there are classification reports for four different models: Basic CNN, Improved CNN, 1D CNN with GRU, and LSTM.

First, the Basic CNN model shows high precision and recall for most emotions. Especially for ”Neutral”, ”Disgust”, and ”Surprised”, the scores are close to or reach 1.00. The overall accuracy of this model is 0.94. The Improved CNN model also performs well in precision and recall for most emotions. It gets a perfect score of 1.00 for ”Neutral” and ”Disgust”. The overall accuracy for this model is slightly better at 0.96. The 1D CNN with GRU model has some weaker recall scores for emotions like ”Calm” and ”Happy”. However, for ”Disgust” and ”Surprised”, the scores are almost perfect. This model has an overall accuracy of 0.94. The LSTM model has balanced precision and recall scores for all emotions. Its overall accuracy is 0.92, which is a bit lower than the other three models.

From figure 3, the validation accuracy for the four models over 150 epochs can be seen. The accuracy for all models fluctuates between 0.85 and 0.95. The Basic CNN and Improved CNN models perform slightly better for most epochs.

In conclusion, all models demonstrate high classification accuracy, especially for certain emotions. However, the Improved CNN model seems to have a slight edge over the others.

V. Evaluation Of the Optimal Model

It has been determined that the improved CNN model, trained on the song subset with MFCC features, achieves the best accuracy. Next, this paper will look more into this model.

The figure 4 presents the model accuracy over epochs for both training and test data. Initially, there is a sharp rise in accuracy for both curves. The training accuracy tends to steadily increase, approaching close to 100% around the 140th epoch. In contrast, the test accuracy, after its initial surge, fluctuates slightly but remains mostly stable from 60 epochs onward, hovering just below 95%. This suggests the model has a commendable fit to the training data. However, the slight divergence between the test and train curves, especially in the latter epochs, may hint at a mild overfitting, where the model

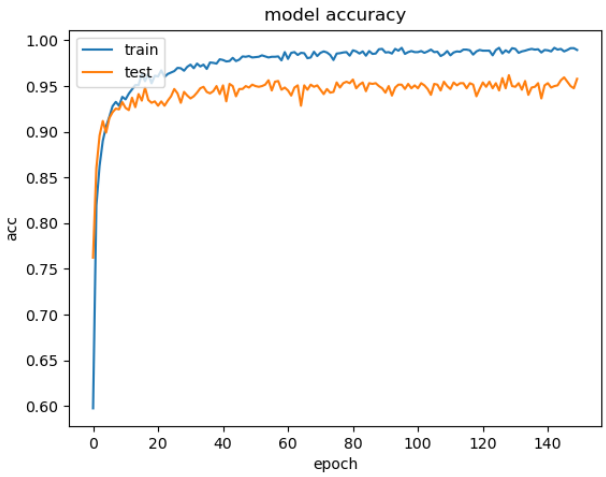

Fig. 4: Accuracy for the Optimal Model

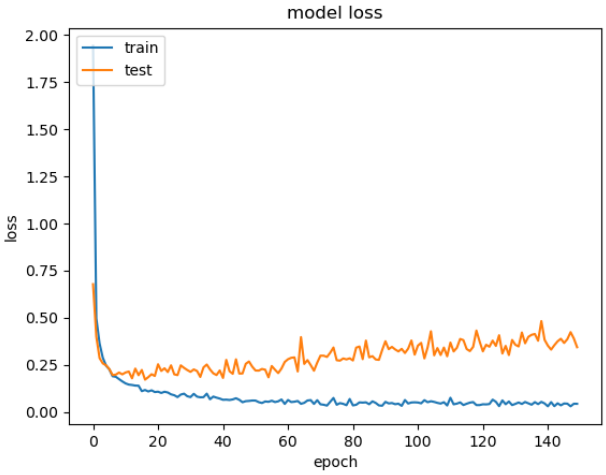

Fig. 5: Loss for the Optimal Model

is overly tailored to the training data but may not perform as optimally on new data.

The figure 5 shows the model loss over epochs for the training and test datasets. It can be seen that the loss on the training set drops sharply at the beginning and stabilizes at a level close to zero as we enter later epochs. The test loss initially follows a similar trend but starts to stabilize around 0.25 after about 60 epochs. While the training loss is almost small, indicating a good fit with the training set, the stabilization of the test loss suggests that the model may have reached its optimization limit.

The confusion matrix from figure 6 shows how a model predicts emotions compared to the true emotions. The main diagonal represents correct predictions. For example, ”neutral” was correctly predicted 165 times. Most emotions, like ”happy” and ”angry,” have high correct predictions. But there are some mistakes. ”Calm” was sometimes mistaken as ”happy,” and ”sad” as ”fearful.” Also, ”fearful” had many

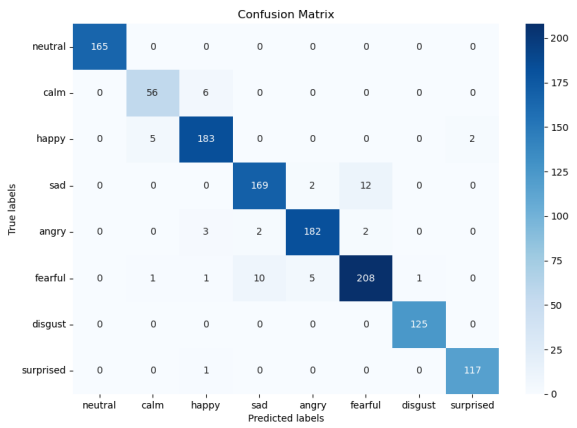

Fig. 6: Confusion Matrix for the Optimal Model

```
classification_report
              precision    recall  f1-score   support

           0       1.00      1.00      1.00       165
           1       0.90      0.90      0.90        62
           2       0.94      0.96      0.95       190
           3       0.93      0.92      0.93       183
           4       0.96      0.96      0.96       189
           5       0.94      0.92      0.93       226
           6       0.99      1.00      1.00       125
           7       0.98      0.99      0.99       118

    accuracy                           0.96      1258
   macro avg       0.96      0.96      0.96      1258
weighted avg       0.96      0.96      0.96      1258
```

Fig. 7: Classification Report for the Optimal Model

wrong predictions, mainly confused with "sad." The model did well for emotions like "neutral," "happy," and "surprised," but can be improved for "fearful" and "calm."

The classification report displays performance metrics for different classes. Classes 0, 6, and 7 have perfect scores of 1.00 in precision, recall, and f1-score. This means they are predicted very well. Classes 1 and 5 have slightly lower scores, with precision and recall around 0.90 to 0.94. The overall model accuracy is high at 0.96, indicating that it correctly predicts 96% of the samples. The macro and weighted averages for precision, recall, and f1-score are also 0.96, showing consistency in performance across classes. This model provides strong results, but there's some room for improvement in certain classes like 1 and 5. The conclusion obtained in Figure 7 is consistent with the conclusion we obtained in the confusion matrix.

VI. Discussion Of Results

The results from the study indicate a high accuracy in emotion prediction using the optimal model. The training accuracy showed a steady increase, nearing 100% by the 150th epoch. On the other hand, the test accuracy, after an initial surge, remained stable from 60 epochs onward, just below 95%. This suggests that the model fits the training data well. However, there is a slight divergence between the test and train curves in the latter epochs, hinting at potential mild overfitting. The confusion matrix revealed that while the model performed well for emotions like "neutral," "happy," and "surprised," there were areas of improvement for emotions like "fearful" and "calm." For instance, "calm" was sometimes misclassified as "happy," and "sad" as "fearful." The classification report further supported these findings, with classes 0, 6, and 7 achieving perfect scores in precision, recall, and f1-score. However, classes 1 and 5 had slightly lower scores, the reason may be that the samples are too single and their MFCC features are too similar. The overall model accuracy was high at 96%, but there's room for enhancement in certain classes.

VII. Conclusion

The study successfully used the model to predict emotions from audio data with an overall accuracy of 96%. The model's performance on some aspects of emotion is commendable, but there are still areas where it could be improved. The results from the confusion matrix and classification reports are consistent, highlighting the model's strengths and areas of potential enhancement. Future work could focus on refining the model to address the observed misclassification issues, or use K-fold cross-validation and stratified K-fold cross-validation to more fully evaluate the model's performance and further improve its accuracy of emotion prediction.

References

[1] M. G. de Pinto, M. Polignano, P. Lops and G. Semeraro, "Emotions Understanding Model from Spoken Language using Deep Neural Networks and Mel-Frequency Cepstral Coefficients," 2020 IEEE Conference on Evolving and Adaptive Intelligent Systems (EAIS), Bari, Italy, 2020, pp. 1-5, doi: 10.1109/EAIS48028.2020.9122698.

[2] T. Hussain, W. Wang, N. Bouaynaya, H. Fathallah-Shaykh and L. Mihaylova, "Deep Learning for Audio Visual Emotion Recognition," 2022 25th International Conference on Information Fusion (FUSION), Linköping, Sweden, 2022, pp. 1-8, doi: 10.23919/FUSION49751.2022.9841342.

[3] F. Andayani, L. B. Theng, M. T. Tsun and C. Chua, "Recognition of Emotion in Speech-related Audio Files with LSTM-Transformer," 2022 5th International Conference on Computing and Informatics (ICCI), New Cairo, Cairo, Egypt, 2022, pp. 087-091, doi: 10.1109/ICCI54321.2022.9756100.

[4] G. Deshmukh, A. Gaonkar, G. Golwalkar and S. Kulkarni, "Speech based Emotion Recognition using Machine Learning," 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 2019, pp. 812-817, doi: 10.1109/ICCMC.2019.8819858.

[5] R. Mittal, S. Vart, P. Shokeen and M. Kumar, "Speech Emotion Recognition," 2022 2nd International Conference on Intelligent Technologies (CONIT), Hubli, India, 2022, pp. 1-6, doi: 10.1109/CONIT55038.2022.9848265.

[6] R. Anusha, P. Subhashini, D. Jyothi, P. Harshitha, J. Sushma and N. Mukesh, "Speech Emotion Recognition using Machine Learning," 2021 5th International Conference on Trends in Electronics and Informatics (ICOEI), Tirunelveli, India, 2021, pp. 1608-1612, doi: 10.1109/ICOEI51242.2021.9453028.

[7] Livingstone SR, Russo FA (2018) The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. PLoS ONE 13(5): e0196391. https://doi.org/10.1371/journal.pone.0196391.

[8] M. G. de Pinto, "Audio Emotion Classification from Multiple Datasets," GitHub, Oct. 22, 2023. https://github.com/marcogdepinto/emotion-classification-from-audio-files (accessed Oct. 28, 2023).