Temple University

CIS 5603
Artificial Intelligence

Professor: Dr. Pei Wang

Project Report: Emotion Detection

Student: Punrong RANY

An assignment submitted in partial fulfillment of the requirement for CIS 5603

Fall 2023

Table of Contents

Abstract

Emotion recognition is a pivotal aspect of human-computer interaction, impacting domains from affective computing to mental health assessment. This project aims to develop a robust emotion recognition system by integrating cutting-edge technologies, specifically facial expression analysis and speech emotion recognition. The comprehensive approach involves the synergistic fusion of outputs from these modalities to enhance accuracy and responsiveness in real-time emotion prediction. The significance lies in the potential to elevate user experiences in applications like virtual assistants, customer service, and educational tools, ultimately advancing human-centric technology interfaces.

The project is structured into three phases: Facial Emotion Detection, Speech Emotion Detection, and the integration of predictions from facial and speech models using the late fusion technique. The experimentation involves deep learning techniques applied to diverse datasets for facial and speech modalities. The facial emotion detection model demonstrates high accuracy of 93% with balanced performance metrics, while the speech emotion detection model achieves an accuracy of 77%, with recommendations for addressing imbalances in specific classes.

The future direction involves bridging performance disparities between facial and speech models for successful integration. The emphasis is on refining the speech model to achieve comparable accuracy and loss metrics with the facial model. Systematic analysis and optimization of the speech model, including hyperparameters and preprocessing techniques, are crucial. Only when both modalities align in performance metrics can the late fusion approach be implemented effectively, ensuring a unified and accurate synthesis of facial and speech-based emotional insights.

Emotion Detection

1. Introduction

Emotion recognition is a critical component of human-computer interaction, with applications spanning from affective computing to mental health assessment. Understanding and accurately identifying human emotions can significantly enhance the capabilities of various systems and services. This project aims to develop an emotion recognition system to achieve a more robust and accurate emotion prediction.

2. Problem Statement

Emotion detection has gained significant attention in recent years due to its potential in various domains, including healthcare, marketing, and entertainment. The problem addressed by this project can be defined as follows:

2.1 Problem Definition

To develop an advanced emotion detection system that can accurately identify and classify a person's emotions in real-time, I am integrating cutting-edge technologies such as facial expression analysis and speech emotion recognition. This comprehensive approach aims to synergistically fuse the outputs of these modalities, enhancing the system's accuracy and responsiveness in capturing the dynamic spectrum of human emotions.

2.2 Why it Matters?

Precise emotion detection plays a pivotal role in elevating human-computer interaction, refining mental health assessments, and enriching user experiences across diverse applications like virtual assistants, customer service, and educational tools. The integration of accurate emotion detection technologies not only enhances the effectiveness of these applications but also contributes to the overall advancement of human-centric technology interfaces

3. Scope

This project plan encompasses the development of an emotion detection system using the deep learning techniques. The project aims to recognize a range of basic emotions, including happiness, sadness, anger, fear, disgust, surprise, and neutral expressions. It will not delve into more complex emotional states or facial analysis for age, gender, or identity recognition. The implementation of this project will be executed in a phased approach, breaking it down into distinct stages, each with its specific objectives and tasks. However, it is important to acknowledge that time constraints may influence the scope of this project.

4. Phases

This project aims to develop an emotion recognition system that leverages multimodal data, combining both facial expressions and speech. The project will be structured around three main phases:

Phase 1: Facial Emotion Detection

The first goal of this project is to create a robust facial expression recognition system. This system will utilize deep learning techniques to analyze facial expressions in real-time, extracting valuable features and predicting the associated emotional state.

Phase 2: Speech Emotion Detection

The second phase focuses on developing a speech emotion recognition system. This system will analyze audio recordings to detect and classify emotional states based on the speaker's voice patterns, pitch, and tone. The performance of the speech emotion recognition model will be enhanced through rigorous training and testing.

Phase 3: Combining Predictions from Facial and Speech Models

In the concluding phase, my objective is to implement the late fusion technique to merge the outputs from both the facial expression and speech emotion recognition models. The successful implementation of late fusion will yield an integrated emotion recognition system proficient in

accurately identifying and categorizing human emotions through the combined analysis of facial expressions and speech.

5. Objectives

The primary objectives of this AI class project are as follows:

a. **Data Collection and Preprocessing**: Collect appropriate datasets comprising facial images and speech samples, annotate them with corresponding emotions, and preprocess the data to prepare it for model training. This involves curating diverse datasets for both facial and speech modalities, annotating the emotional content, and applying necessary preprocessing steps to optimize the data for effective model training.
b. **Model Development**: Develop deep learning models capable of recognizing and classifying human emotions based on facial expressions and speech.
c. **Model Evaluation**: Evaluate the performance of the models in terms of accuracy, precision, and other pertinent metrics to gain insights into their effectiveness. This assessment encompasses multiple models, providing a comprehensive understanding of their individual and collective capabilities.
d. **Real-time Emotion Detection**: Implement real-time facial and speech emotion detection and assess its efficiency.

6. Dataset

For facial emotion detection, I am going to utilize two datasets such as FER-2013 by Manas Sambare and Facial Emotion Expressions by Samaneh Eslamifar. These two datasets comprise grayscale facial images with dimensions of 48x48 pixels. These images have undergone automated alignment to ensure that the face is approximately centered and occupies a consistent portion of the image.

To build a robust speech emotion detection system, I have used four datasets. The RAVDESS Emotional Speech Audio dataset by Steven R. Livingstone includes 1440 files with emotions expressed by 24 actors. The Toronto emotional speech set, exclusively featuring female speakers,

comprises 200 target words across seven emotions. Crowd Crowd Sourced Emotional Multimodal Actors Dataset, a diverse dataset with 7,442 clips, spans various actors, sentences, emotions, and intensity levels. Surrey Audio-Visual Expressed Emotion focuses on male speakers, providing high-quality audio recordings for seven emotions.

7. Literature Review

7.1 Facial Emotion Detection

Facial Emotion Recognition through deep learning has gained significance in various fields, including safety and healthcare. Researchers aim to decode facial expressions automatically for improved computer-based predictions. Deep learning has played a pivotal role in advancing Facial Emotion Recognition, leading to the exploration of diverse architectural configurations.

Mellouk and Handouzi (2020) conducted a comprehensive review of recent Facial Emotion Recognition works using deep learning. Their analysis underscored the significance of contributions, architectures, and databases, with the aim of guiding researchers to enhance facial emotion recognition.

Huang et al. (2023) conducted a study using deep neural networks, focusing on critical facial features for facial emotion recognition. They found that features around the nose and mouth are key landmarks for neural networks, enhancing Facial Emotion Recognition accuracy. Cross-database validations showed that models pretrained on one dataset and fine-tuned on another improved accuracy.

In conclusion, Facial Emotion Recognition via deep learning is a growing field with applications in safety and healthcare. Recent research emphasizes the importance of specific facial features and cross-database model transfer, offering insights to improve Facial Emotion Recognition systems. Collaboration and further research in this domain hold promise for future advancements.

7.2 Speech Emotion Detection

Over the past decade, Speech Emotion Recognition has emerged as a critical component of Human-Computer Interaction and advanced speech processing systems. Speech Emotion

Recognition aims to identify and classify emotions in speech, contributing to enhanced user experiences.

Aouani and Ben Ayed (2020) proposed a two-stage Speech Emotion Recognition approach, emphasizing feature extraction and classification. They introduced a 42-dimensional vector of audio features, including MFCC coefficients, ZCR, HNR, and TEO, along with the innovative use of Auto-Encoder. SVM was employed for classification, with experiments conducted on the RML dataset.

Wani et al. (2021) conducted a comprehensive review of Speech Emotion Recognition systems, highlighting the interdisciplinary nature of this field. Quantitative and qualitative disparities in human and machine emotion recognition pose challenges. This review identified research gaps and the need for cross-disciplinary collaboration to advance Speech Emotion Recognition systems.

In summary, Speech Emotion Recognition continues to evolve, with a focus on feature extraction and classification to enhance emotion detection in speech. Interdisciplinary cooperation is vital for addressing research gaps and advancing the field.

7.3 Emotion Detection by Facial Expression and Speech

In the 2009 paper by Emerich, Lupu, and Apatean, a bimodal emotion recognition system is introduced, combining facial expressions and speech signals. Using a dataset featuring six emotions and ten subjects, various classifiers, including Support Vector Machine, Naive Bayes, and K-Nearest Neighbor, were employed.

Crucially, the study presented two fusion approaches: feature level and match score level fusion. This is in line with the broader trend of integrating data from different modalities to enhance emotion recognition system performance. Comparative analysis showed that fusion-based techniques indeed improve performance and robustness. Notably, feature-level fusion outperformed score-level fusion, emphasizing the significance of early integration of multimodal features.

This research underscores the increasing importance of multimodal approaches in emotion recognition, offering valuable techniques for enhancing human-computer interaction and our understanding of emotions. As technology advances, the field will continue to evolve, yielding even more sophisticated systems.

8. Experiments

8.1 Facial Emotion Detection

8.1.1 Methodology

8.1.1.1 Dataset

For this study, we utilize two primary datasets mentioned above. After merging these datasets, the resulting dataset employed in our experiments comprises 57,530 images for training and 14,244 images for testing.

8.1.1.2 Model Architecture

a. Overview

The model architecture is designed with a thoughtful configuration to capture intricate patterns in facial expressions. It consists of five convolutional layers, one flatten layer, three fully connected layers, and one output layer with softmax activation for multi-class classification. Each convolutional layer is augmented with Batch Normalization, ReLU Activation, MaxPooling, and Dropout layers. Similarly, each fully connected layer incorporates L2 regularization, Batch Normalization, ReLU Activation, and Dropout layers. The model is compiled with the Adam optimizer and categorical crossentropy loss, and its summary reveals the layer-wise configuration, emphasizing the thoughtful design to balance complexity and generalization.

b. Training Optimization Techniques

To ensure the model's robustness and prevent overfitting, several optimization techniques are employed:

- **Early Stop**: Implemented to halt training when overfitting is detected, thereby preventing unnecessary resource consumption.

- **Model Saving Callback**: This callback saves the model with the best performance during training, ensuring that the most optimal configuration is retained.
- **Pre-defined Reduce Learning Rate Callback**: Adaptive learning rate adjustment, triggered when improvements in the chosen metric become less frequent, aids in fine-tuning the model.
- **Class Weight Computation**: Recognizing the imbalances in the training data, a class imbalance-aware technique is applied to compute class weights, addressing issues related to uneven representation of emotions.

These carefully chosen strategies collectively contribute to the comprehensive training and optimization of our facial emotion recognition model, enhancing its performance and generalization capabilities.

8.1.2 Result and Discussion

The neural network underwent a 200-epoch training process, revealing a consistent decrease in both training and validation losses over time. The model's accuracy steadily improved throughout the training period, culminating in a noteworthy accuracy of 96.96% on the training set and 92.89% on the validation set by the final epoch. This indicates successful learning and generalization capabilities, showcasing the model's effectiveness in capturing patterns within the dataset.

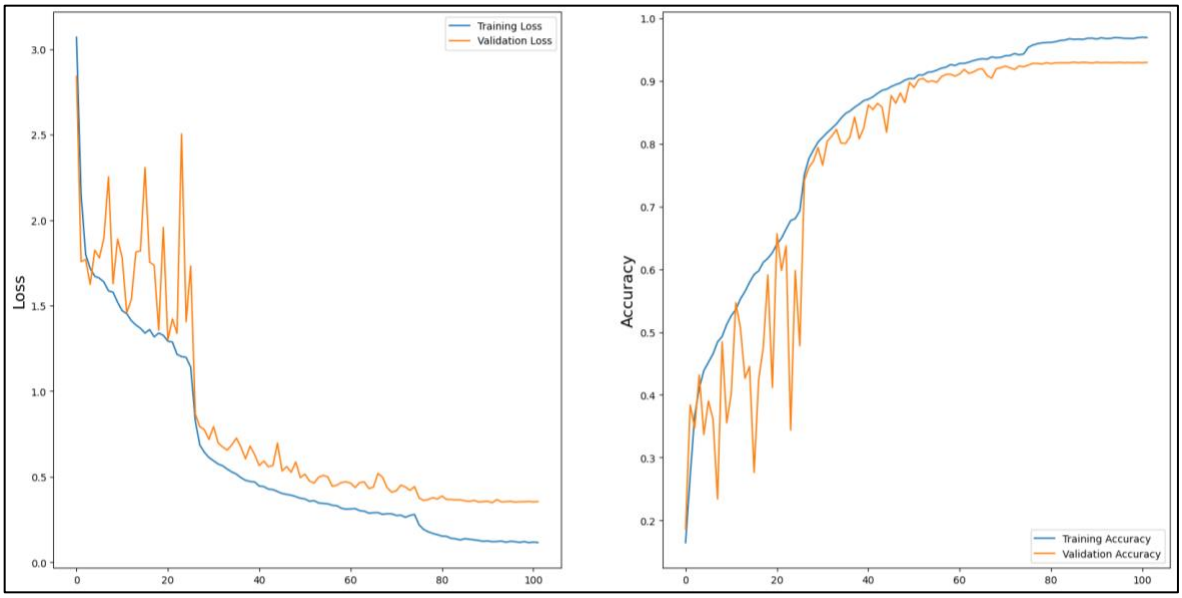

Figure 1. Training and Validation Metrics Over Epochs for Facial Emotion Detection Model

The classification report reveals the model's consistent precision and recall, affirming its accuracy in emotion classification. With a macro-average F1-score of 0.92, the model exhibits a well-balanced performance across diverse emotions, minimizing bias. Impressively, the model achieves an overall accuracy of 93%, underscoring its proficiency in accurately predicting emotional states. The robustness of the model is evident in strong weighted average metrics, indicating consistent performance across the entire dataset and highlighting its generalization capabilities. However, the report identifies an imbalance in the "disgust" class (support: 222). To address this issue, it is recommended to employ techniques such as data augmentation or resampling, ensuring a more balanced representation of the "disgust" class and further enhancing the model's overall performance and reliability.

	precision	recall	f1-score	support
angry	0.92	0.91	0.91	1918
disgust	0.95	0.94	0.94	222
fear	0.92	0.89	0.91	2042
happy	0.97	0.96	0.97	3599
neutral	0.91	0.93	0.92	2449
sad	0.89	0.90	0.90	2386
surprise	0.94	0.96	0.95	1628
accuracy			0.93	14244
macro avg	0.93	0.93	0.93	14244
weighted avg	0.93	0.93	0.93	14244

Figure 2. Classification Report for Facial Emotion Detection Model with Precision, Recall, and Recommendations

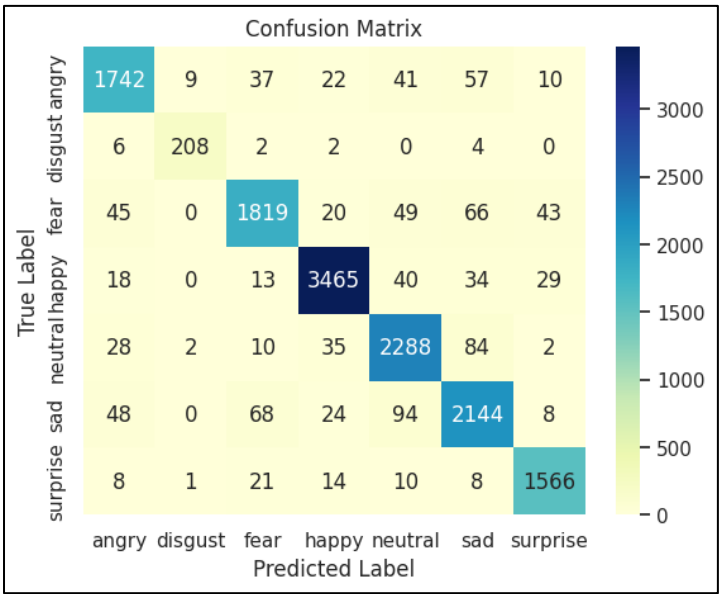

Figure 3. Confusion Matrix for Facial Emotion Detection Model

8.1.3 Real-time webcam

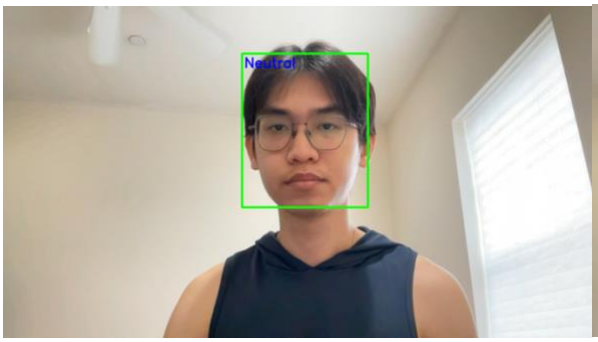

Figure 4. Neutral Facial Expression

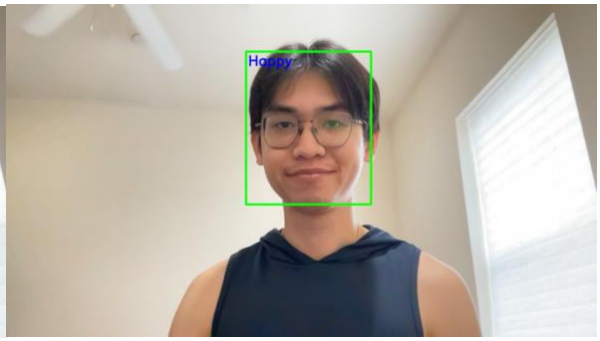

Figure 5. Happy Facial Expression

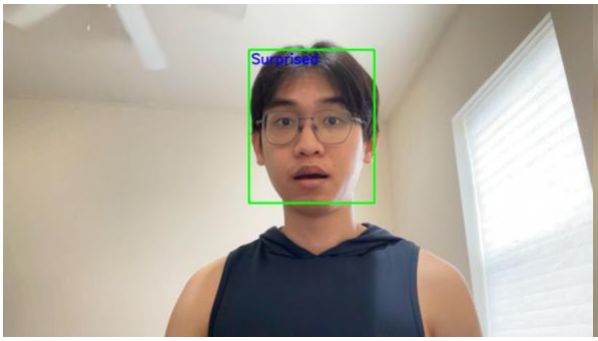

Figure 6. Surprised Facial Expression

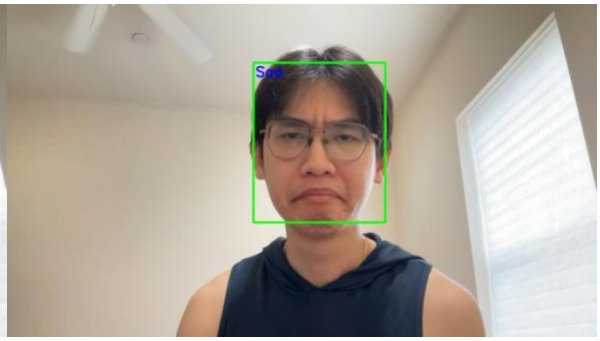

Figure 7. Sad Facial Expression

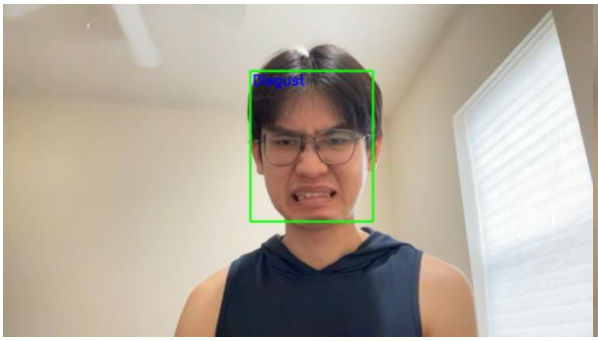

Figure 8. Disgust Facial Expression

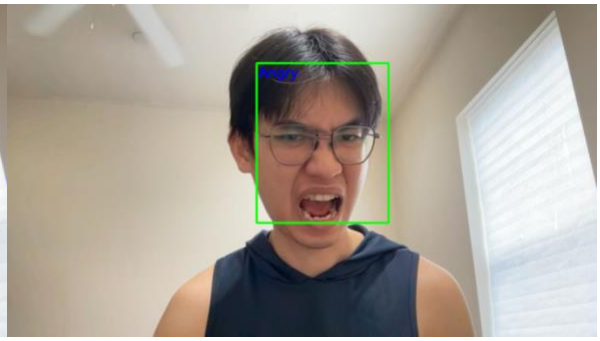

Figure 9. Angry Facial Expression

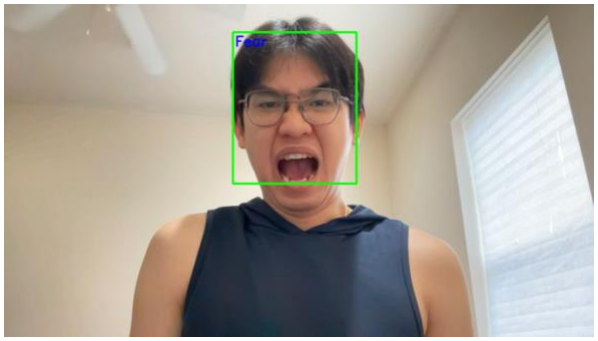

Figure 10. Fear Facial Expression

8.2 Speech Emotion Detection

8.2.1 Methodology

8.2.1.1 Dataset

In this study, I am utilizing four datasets shown above by concatenating them all for speech emotion detection. The resulting dataset employed in our experiments comprises audio samples for training and testing.

8.2.1.2 Methodology

a. Overview

The model architecture is designed for speech emotion detection, incorporating Convolutional Neural Network layers to capture intricate patterns in audio features. The model consists of four convolutional layers with increasing filter sizes, each followed by Batch Normalization, ReLU Activation, MaxPooling, and Dropout layers. Additionally, the model includes three fully connected layers with L2 regularization, Batch Normalization, ReLU Activation, and Dropout layers to enhance feature learning and prevent overfitting. The output layer utilizes softmax activation for multi-class classification, producing probabilities for seven emotion classes. The model is compiled with the Adam optimizer and categorical crossentropy loss, and its summary reveals the layer-wise configuration, emphasizing the thoughtful design to balance complexity and generalization.

b. Training Optimization Techniques

To enhance the robustness and prevent overfitting of the speech emotion detection model, a suite of training optimization techniques is strategically implemented:

- **Audio Data Augmentation Techniques**: Various audio data augmentation techniques, including noise injection, time stretching, time shifting, and pitch modulation, are employed. These techniques contribute to the model's ability to generalize by exposing it to diverse variations in the training data, ultimately improving its performance on unseen examples.
- **Model Saving Callback**: This callback saves the model with the best performance during training, ensuring that the most optimal configuration is retained.

- **Pre-defined Reduce Learning Rate Callback**: Adaptive learning rate adjustment, triggered when improvements in the chosen metric become less frequent, aids in fine-tuning the model.

These chosen strategies collectively contribute to the comprehensive training and optimization of our speech emotion detection model, enhancing its resilience, performance, and generalization capabilities.

8.2.2 Result and Discussion

The neural network underwent a 200-epoch training process, revealing a consistent decrease in both training and validation losses over time. The model's accuracy steadily improved throughout the training period, culminating in a noteworthy accuracy of 76.05% on the training set and 73.17% on the validation set by the final epoch. This indicates successful learning and generalization capabilities, showcasing the model's effectiveness in capturing patterns within the dataset.

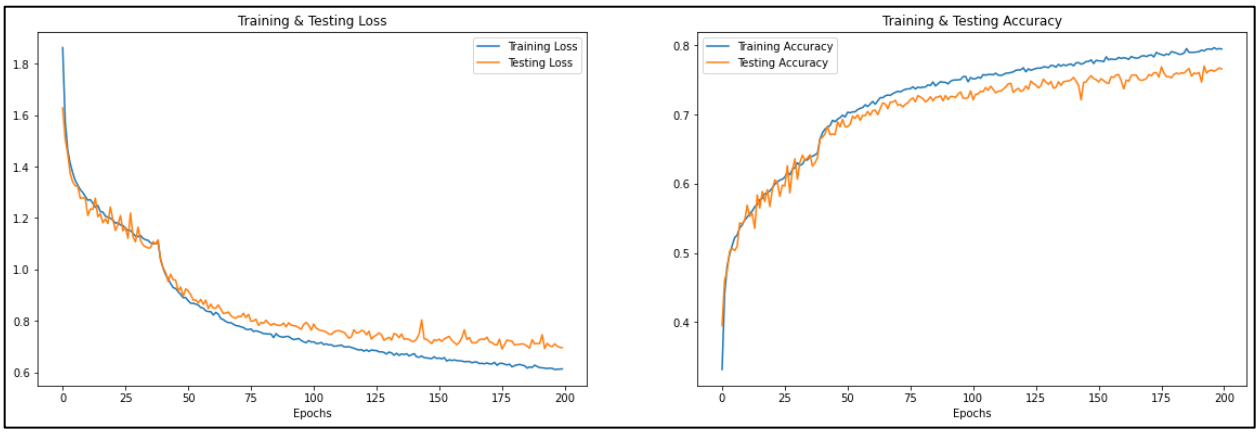

Figure 11. Training and Validation Metrics Over Epochs for Speech Emotion Detection Model

The classification report underscores the model's performance in emotion classification across various categories. Notably, the model demonstrates a commendable precision-recall balance, affirming its accuracy. The macro-average F1-score of 0.78 reflects a well-balanced performance, minimizing bias and showcasing the model's proficiency in capturing diverse emotions.

Impressively, the overall accuracy of 77% attests to the model's ability to predict emotional states accurately. The weighted average metrics further support its robustness, indicating consistent performance across the entire dataset and highlighting its generalization capabilities.

However, a critical observation arises from an imbalance in the "disgust" class, with a precision of 0.67 and recall of 0.73 (support: 1992). To address this issue effectively, it is recommended to implement strategies such as data augmentation or resampling. These techniques can help achieve a more balanced representation of the "disgust" class, ultimately enhancing the model's overall performance and reliability.

	precision	recall	f1-score	support
angry	0.92	0.86	0.89	1892
disgust	0.67	0.73	0.70	1992
fear	0.84	0.66	0.74	1860
happy	0.77	0.79	0.78	1908
neutral	0.63	0.80	0.71	1699
sad	0.77	0.70	0.73	1947
surprise	0.92	0.96	0.94	672
accuracy			0.77	11970
macro avg	0.79	0.78	0.78	11970
weighted avg	0.78	0.77	0.77	11970

Figure 12. Classification Report for Speech Emotion Detection Model with Precision, Recall, and Recommendations

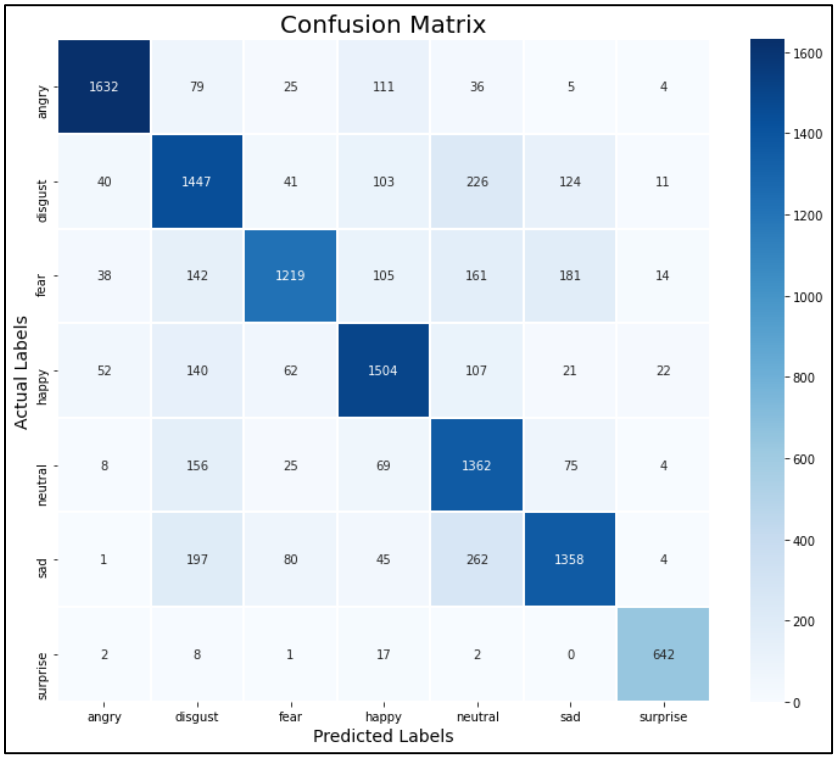

Figure 13. Confusion Matrix for Speech Emotion Detection Model

9. Future Directions

For phase 3, the integration of facial and speech models poses a challenge due to significant disparities in accuracy and loss between the two modalities. Addressing this discrepancy is pivotal to achieving a cohesive and well-balanced integrated model.

It is imperative to bridge the existing disparities in performance metrics between the two modalities, aiming for a convergence where both models demonstrate comparable accuracy levels and loss values. This parity is essential for the seamless fusion of facial and speech data, ensuring a unified and robust emotion recognition system that accurately captures emotional states across diverse modalities. Efforts should be concentrated on refining the speech model to attain a balance that mirrors the well-established accuracy and loss benchmarks set by the facial model.

To enhance the performance of the speech model, a systematic analysis of its training process should be conducted. This involves fine-tuning hyperparameters, optimizing network architecture, and experimenting with various preprocessing techniques tailored to the intricacies of speech data.

Only when both modalities exhibit consistent accuracy and loss metrics can the late fusion approach be effectively implemented, ensuring a harmonized and accurate synthesis of facial and speech-based emotional insights.

Reference

Aouani, H., & Ben Ayed, Y. (2020). Speech Emotion Recognition with deep learning. Procedia Computer Science, 176, 251-260. https://doi.org/10.1016/j.procs.2020.08.027

Emerich, S., Lupu, E., & Apatean, A. (2009). Emotions recognition by speech and facial expressions analysis. In 17th European Signal Processing Conference (EUSIPCO 2009). https://www.eurasip.org/Proceedings/Eusipco/Eusipco2009/contents/papers/1569192455.pdf

Huang, ZY., Chiang, CC., Chen, JH. et al. A study on computer vision for facial emotion recognition. Sci Rep 13, 8425 (2023). https://doi.org/10.1038/s41598-023-35446-4

Mellouk, W., & Handouzi, W. (2020). Facial emotion recognition using deep learning: review and insights. Procedia Computer Science, 175, 689-694. https://doi.org/10.1016/j.procs.2020.07.101

T. M. Wani, T. S. Gunawan, S. A. A. Qadri, M. Kartiwi and E. Ambikairajah, "A Comprehensive Review of Speech Emotion Recognition Systems," in IEEE Access, vol. 9, pp. 47795-47814, 2021, doi: 10.1109/ACCESS.2021.3068045

Dataset

Facial Emotion Expression Datasets

- FER-2013 by Manas Sambare: https://www.kaggle.com/datasets/msambare/fer2013
- Facial Emotion Expressions by Samaneh Eslamifar: https://www.kaggle.com/datasets/samaneheslamifar/facial-emotion-expressions

Speech Emotion Expression Datasets

- RAVDESS Emotional Speech Audio dataset by Steven R. Livingstone: https://www.kaggle.com/datasets/uwrfkaggler/ravdess-emotional-speech-audio
- Toronto emotional speech set (TESS): https://www.kaggle.com/datasets/ejlok1/toronto-emotional-speech-set-tess
- Crowd Sourced Emotional Multimodal Actors Dataset (CREMA-D): https://www.kaggle.com/datasets/ejlok1/cremad
- Surrey Audio-Visual Expressed Emotion (SAVEE): https://www.kaggle.com/datasets/ejlok1/surrey-audiovisual-expressed-emotion-savee