[1]

Speech Emotion Recognition using CRNN

Adam Kanoun*, David He*, Jai Dey*, Sizhe Fan*
University of Toronto
3359 Mississauga Rd, Mississauga, ON L5L 1C6, Canada
adam.kanoun@mail.utoronto.ca
davidhe.he@mail.utoronto.ca
jai.dey@mail.utoronto.ca
sizhe.fan@mail.utoronto.ca

Abstract

Our project focuses on building a deep learning model to classify human speech into emotions like happiness, sadness, anger, and surprise. Models capable of detecting human emotions would enable use cases such as virtual assistants or music apps that adjust their responses based on how a person feels. To achieve this, we propose to use a type of deep learning model called a Convolutional Recurrent Neural Network. The idea is that our CRNN combines layers that pick up important sound features with layers that track changes over time, making it a good fit for picking up patterns in speech. Firstly, the plan is to apply data preprocessing techniques like removing background noise and converting the audio to visual spectrograms. Then a dynamic learning rate scheduler will be implemented. This essentially allows us to adjust the learning rate of our CRNN during the training process, which optimizes our training speed. Lastly, some extra techniques we plan to use to help generalize speech patterns are pitch shifting and time stretching. Pitch shifting changes the tone of the sound, making it higher or lower. While time stretching changes how fast or slow the sound is but keeps the tone untouched.

1 Introduction

Every day, millions of people interact with digital devices that can do amazing things - such as playing music, answering questions, or helping us stay organized. Yet although these technologies play a huge part in our lives, they are unable to understand how we humans feel.

For this reason, we strive to develop a deep learning model that can classify emotions in human speech. This would allow us to have more enhanced human-machine interactions. For example, an online music playlist could change its song choice by adapting to the user's emotions in real-time. Another example is that virtual assistants could respond to users based on their emotional state. Being able to offer mental health support to people at their lowest point can potentially save lives and prevent disasters such as school shootings or suicides.

Moreover, as it currently stands, humans are inevitably becoming more reliant on digital interfaces within each passing year. For this reason, the ability to detect emotions is essential for the future of humanity if we want to create more enriched lifestyles for the majority of the populace.

To achieve this, our general approach is to first preprocess our data by removing background noise and converting the audio to visual spectrograms. Then, our approach consists of utilising a CRNN to further analyze audio data. This is because a CRNN allows us to automate the feature extraction

[1] Github Repository: https://github.com/RuthlessRu/vigilant-fishstick/tree/main

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

process, and it can also easily handle the complexities of temporal audio patterns due to its recurrent layers. CRNNs have also demonstrated strong performance in related audio recognition tasks, making them very well-suited for this project. Specifically, we chose a female audio dataset to better capture the nuances of speech patterns for this demographic, acknowledging potential differences in vocal characteristics and emotional expression between genders.

2 Background and Related Works

Speech signals are the natural medium of human communication that convey information about the speaker's message, perspective and identity as well as their emotions. Speech Emotion Recognition (SER) is a classification problem that aims to infer the emotional state of a speaker from speech signals. Although recognizing emotions is easy for humans, it is a challenging task for machines. This is because speech itself is a complex signal that differs from person to person and can change depending on context and intonation [?] [?]. In recent years, there have been successful attempts at using Deep Neural Networks (DNN) and Convolutional Neural Networks (CNN) to recognize emotions from speech [?] [?]. These attempts have also stressed the importance of finding meaningful features through feature extraction in SER.

2.1 Feature Extraction for SER

Emotions in speech can be expressed through subtle variations in pitch, tone, energy, and timing. The goal of this step is to find features that can be extracted from input audio to differentiate between different emotions. Previous works have made use of a combination of acoustic features, which include prosodic and spectral features, for SER classification. Language information may be used as well but is often paired with acoustic features [?] [?].

Prosodic features are features that are related to pitch, loudness, rhythm, tempo, and intonation, which give clues about how something is said rather than what is said. These features provide a way to differentiate emotions of different arousal (intensity) levels, such as happy and sad. However, they can't differentiate between emotions of the same arousal level but different valence (positivity), such as happy and angry. Spectral features are features that capture the frequency of speech signals. The most common of these features are Mel-Frequency Cepstral Coefficients (MFCCs) [?]. Although these handcrafted features seem meaningful, they might not be enough for SER [?]. In this case, deep learning can be trained on Mel-spectrograms.

A Mel-spectrogram is a feature that provides a way to visually represent changes in loundness and frequency in a speech signal over time. To provide useful context, the Mel part of the name comes from scaling frequency in a way that matches the human auditory perception. The spectrogram part means that different parts of a speech signal input are mapped from a time domain to a frequency domain using fast Fourier transform before being stacked together. The result of stacking is then a graph that maps time-frequency to loudness [?].

2.2 CRNN for SER

When it comes to SER, deep learning models can provide more desirable performance compared to traditional machine learning models. Audio data is often proccessed into spectrograms, which are 2-dimensional images, before being fed as input.

Since the signal processing task has transformed into an image processing task, Convolutional Neural Networks (CNN) are reasonable choices for feature extraction and selection. The feature extraction of low-level descriptors (LLD) can be done with convolution functions while feature selection can be done with pooling layers such as max-pooling.

Likewise, Recurrent Neural Networks (RNN) are suitable for learning data that is sequential in nature, such as audio data, with its memory mechanism. However, RNNs don't have good gradient flow, so Long Short-Term Memory (LSTM) networks are often used instead [?].

2.3 Attention Mechanism

Not all features of the speech signal might be relevant in recognizing emotions [**?**]. For example, a speaker might sound neutral at the beginning and ending of a sentence but sound angry in the middle. The classifier should focus on the features that occur in the middle of the sequence. This leads to using attention mechanisms to improve the performance of LSTM networks by putting more emphasis on certain parts of the input sequence.

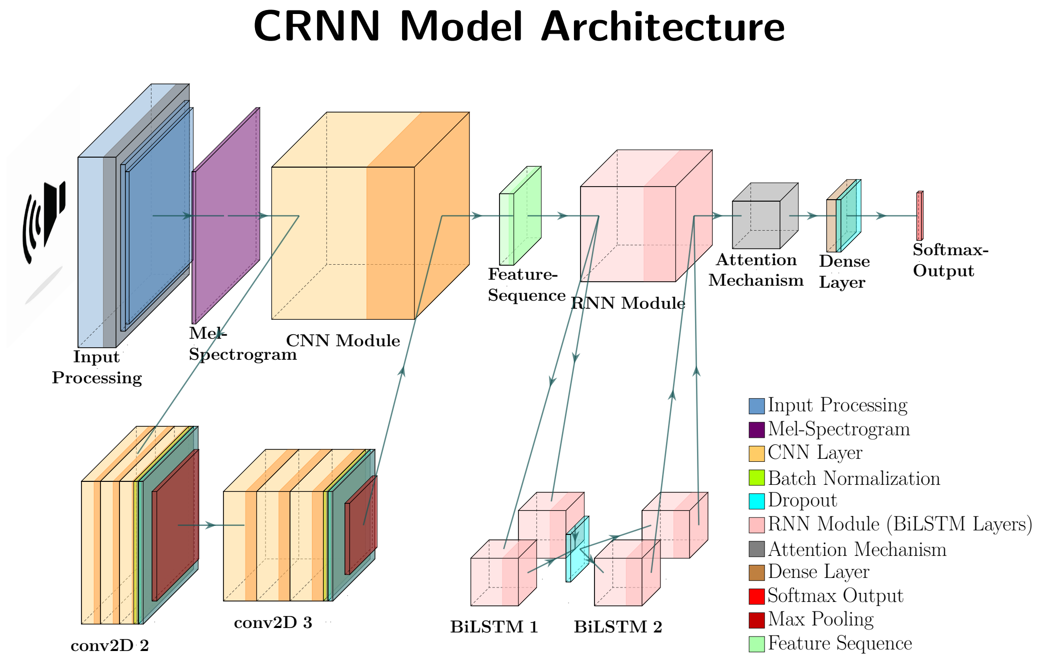

Figure 1: First, audio inputs are processed into Mel-spectrograms. Mel-spectrograms go through a series of convolution, batch normalization, dropout, and max-pooling in the CNN module. The result is then a feature sequence that goes through two Bidirectional LSTM layers in the RNN module with dropout in between. An attention mechanism is applied afterwards followed by a fully connected layer with dropout. Finally, softmax is applied.

3 Data

The Toronto Emotional Speech Set (TESS), Crowd-Sourced Emotional Multimodal Actors Dataset (CREMA-D), and Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) are well-established datasets designed for emotion recognition in speech. We focus on six distinct emotions across all datasets: neutral, happy, sad, angry, fear, and disgust. TESS contains audio samples from two female actors (a younger and older adult), CREMA-D consists of 7,442 clips from 91 actors (48 male, 43 female), and RAVDESS includes recordings from 24 professional actors (12 female, 12 male). To maintain consistency, we only utilize recordings from female speakers across all datasets. Then, we combine the TESS, CREMA-D, and RAVDESS datasets by splitting each dataset evenly into their respective train, validation, and test sets. Lastly, we merge each set of the same type (e.g., train sets, validation sets, test sets) together to create a unified dataset for training, validation, and testing purposes. This approach allows us to train on a diverse range of speakers while maintaining balanced representation across emotional classes, which is crucial for developing unbiased models.

Additionally, to enhance the robustness of the model, we apply audio-specific data augmentation. Augmentations such as random pitch shifting and time-stretching effectively expand our dataset by introducing slight variability in existing data, in turn making our model less prone to overfitting [**?**]. These protocols ensure model robustness, aiding us in our goal of building a truly useful and reliable voice-based emotion classifier.

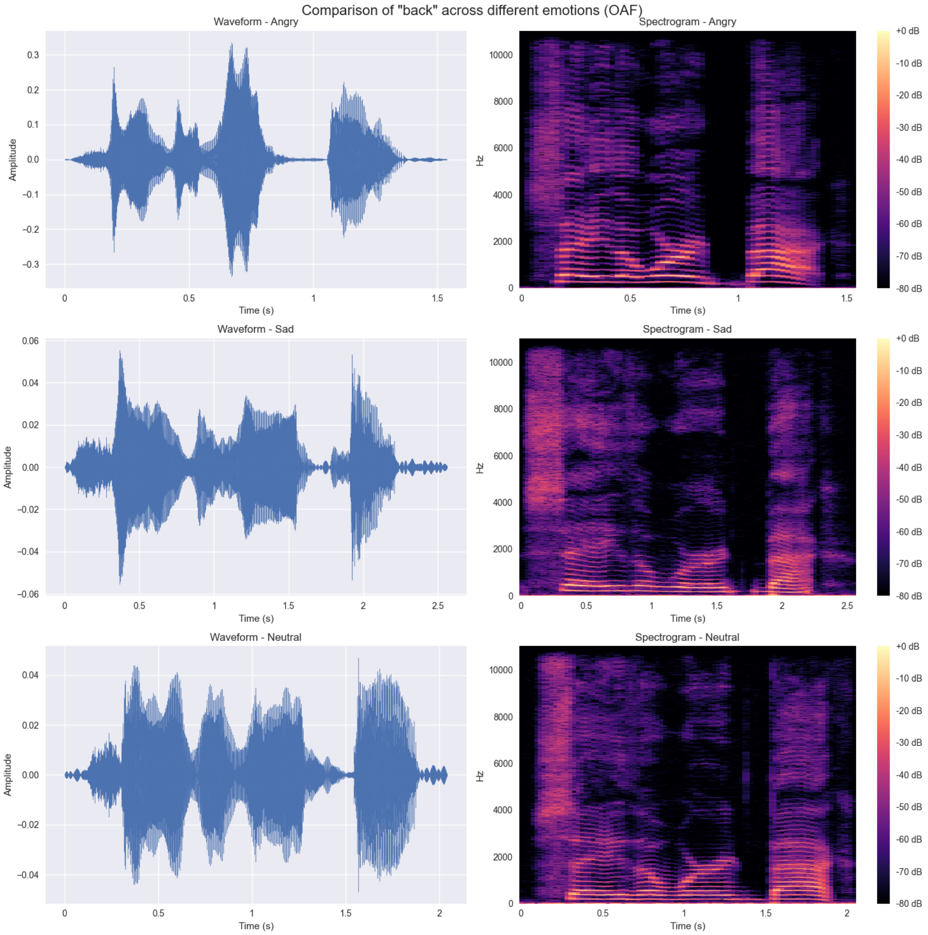

Figure 2: Comparison of waveforms and spectrograms of the word "back" spoken with angry, sad, and neutral emotions by the older adult female speaker (OAF). Left column shows the time-domain waveforms, the right column shows the corresponding spectrograms with intensity in dB.

Furthormore, a key part of SER is extracting useful features from speech signals. The images shown in Figure 2 reveal several key patterns in how emotions affect speech characteristics. The waveforms show clear differences in amplitude, with angry speech having the highest amplitude (±0.3), while sad and neutral speech show much lower amplitudes (±0.06 and ±0.04 respectively). In the spectrograms, we can observe that angry speech displays more intense energy (shown by brighter colors) especially in the 0-5 kHz range, while sad and neutral expressions show more diffused energy patterns.

The temporal characteristics also vary notably across emotions. As seen in the time axes, the sad utterance is the longest (around 2.5 seconds), while the angry utterance is more compact (1.5 seconds), and the neutral utterance takes a moderate duration (2 seconds). Looking at the spectrograms' frequency distribution, angry speech shows distinct, high-intensity patterns (visible as bright red-purple regions), whereas sad and neutral expressions share similar, more subtle spectral patterns. These acoustic similarities between certain emotion pairs, combined with varying temporal patterns, suggest our model will need to consider both timing and frequency characteristics to effectively classify emotions.

4 Ethical Considerations

In utilizing the Toronto Emotional Speech Set (TESS), we acknowledged several important ethical considerations regarding privacy, bias, and potential applications. While the dataset includes proper consent from the two female actors and is publicly available, we recognized that voice data is inherently personal and requires careful handling. The dataset has notable representation limitations: it only includes female voices, is limited to two age groups, and contains only English recordings from Canadian speakers, potentially affecting our model's generalization capabilities. Additionally, since the emotions are performed rather than naturally occurring, this may limit the applicability of our results in the real world. We also acknowledged that emotion recognition technology could potentially be misused for unauthorized surveillance or biased decision-making. To address these concerns, we committed to the following:

1. Bias Mitigation: Recognizing the dataset's demographic limitations, we aimed to use data augmentation techniques such as noise injection, time stretching, and pitch to simulate a broader range of voices. However, we were also transparent about the limitations that these techniques imposed, as simulated diversity cannot truly replace real-world diversity in voice samples.
2. Usage and Transparency: We committed to using the data solely for emotion classification research, and openly addressed the challenges of using acted emotion data and single-gender samples.

5 Work Division

Each team member met every week Monday 9-10 pm to discuss the progress and next steps of the final project. At each meeting, individual members were assigned a task that they are responsible for completing by the end of the week. Collaboration was done asynchronously through an online video and messaging platform with writing done on a shared document.

5.1 Project Proposal

The project proposal responsibilities were equally divided. Jai wrote on the abstract and introduction. David worked on the background and related works. Adam designed the model architecture figure and wrote the architecture section. Fan wrote the data section and collaborated with Adam on the ethical considerations section.

5.2 Final Report

The final report was done in two parts: model buiding and performance analysis. Model building which include data processing, augmentation, training, and testing is done by Adam. The performance analysis which involve results, discussion and limitations of the model as well as report revisions are done by David and Jai.

6 Model Architecture

For SER, we use a Convolutional Recurrent Neural Network (CRNN) architecture that provides both feature extraction and selection capabilities from CNN and sequential learning from RNN.

6.1 Input Processing

In order to ensure uniformity across our datasets, we apply a standardized preprocessing pipeline to the raw audio signals. First, each sound bite is resampled to 16kHz and then normalized to a consistent amplitude. Second, silence segments are removed to minimize irrelevant input (noise) [**?**], and segmented to a fixed length of 2.5 seconds; shorter samples are padded while longer samples are trimmed. Finally, each processed sample is converted into a Mel-spectrogram with 64 mel bands and a maximum frequency of 8kHz. In effect, this transformation reduces the dimensionality of the audio data while preserving essential information about the temporal and spectral characteristics of the speech [**?**].

6.2 Convolutional Layers

The CRNN architecture is initiated by two convolutional layers designed to extract features from the spectrogram input. The first layer transforms the input using 32 filters with 3×3 kernels to capture micro-level emotional indicators and basic acoustic patterns. The second layer expands to 64 filters while maintaining the 3×3 kernel size to capture broader, higher-level abstractions, such as intonation patterns or general timbral characteristics of certain emotions. Each convolution is followed by a max-pooling layer which reduces the spatial dimensions of the feature maps. Moreover, batch normalization and dropout (0.3) is applied after each convolutional layer to stabilize learning and prevent overfitting. These convolutional layers help the model "learn effective salient features for SER and show excellent performances on several benchmark datasets" [?].

6.3 Recurrent Layers

Following the convolutional layers, the model incorporates two bidirectional Long Short-Term Memory (BiLSTM) layers. The first BiLSTM processes the reshaped features with 128 hidden units in each direction, while the second uses 64 hidden units. Given that the audio data is relatively short and sequential, BiLSTMs are a practical choice as they capture dependencies across time steps, and allow for a more nuanced understanding of emotional expression. Additionally, the bidirectional design of the model helps capture the full emotional arc of an utterance by analyzing past and future time frames [?]. Thus, the RNN module serves as a temporal feature extractor [?], supporting the objective of emotion identification over the course of the audio sample.

6.4 Attention Mechanism

To help the model focus on the most salient temporal features for emotion recognition, an attention mechanism is incorporated after the BiLSTM layers [?], [?]. More precisely, the attention module calculates a weighted sum of the hidden states from the BiLSTM, where the weights are determined by both the current hidden state and the overall sequence of hidden states [?]. This can be exceptionally useful since it allows the model to dynamically weigh the importance of different time steps, emphasizing parts of the audio where emotional cues are the strongest, thus possibly improving model performance.

6.5 Dense Layers and Output Layer

Finally, after both the CNN and RNN modules, the resulting features are passed through two fully-connected dense layers. The first dense layer reduces the dimension to 64 units with ReLU activation, followed by dropout (0.3) for regularization. This integration layer is crucial since it combines our high-level temporal and spatial features into a single, unified representation [?]. The final layer outputs logits for our six emotion classes (neutral, happy, sad, angry, fear, and disgust), which are transformed into probabilities using softmax.

6.6 Training Protocol

The training strategy is specifically designed around the nuanced nature of emotional speech patterns. The model employs categorical cross-entropy loss (standard for multi-class classification), with the Adam optimizer (learning rate of 0.001) for its ability to handle the inherent noise and variability in emotional expression [?]. We employ a ReduceLROnPlateau scheduler that reduces the learning rate by half when validation loss plateaus, with a patience of 3 epochs. Training continues for 30 epochs with a batch size of 32.

7 Limitations

The greatest limitation with the proposed approach to SER is the lack of a large and diverse dataset to train on. For instance, the TESS dataset contains 2800 audio data given by two female actors. Although the CREMA-D dataset does provide more diversity with with more 91 actors, it isn't enough for the task of general SER. The use of audio data augmentation could solve this issue somewhat, however, the real world has a variety of vocal nuances that are difficult to replicate with augmentation. A more ideal solution would be to include a more diverse range actors. However, gathering such data and handling its computational demands is challenging. Even if we were to overcome these challenges, the model would still be limited to adult English speakers.

Moreover, the model's complexity increases overfitting risks, especially with small datasets. While dropout layers and early stopping can help, simpler models, like a CNN with attention mechanisms, may be better suited for smaller datasets like TESS.

8 Results

We chose accuracy as our primary performance metric to evaluate our emotion classification model. This is because accuracy provides a straightforward measure of the proportion of correctly classified samples. Moreover, with a balanced dataset, accuracy alone is able to reflect the model's classification capability without requiring additional metrics like precision or recall.

8.1 Baseline Performance

A ResNet-based CNN from Torchvision was chosen as the baseline model, since it performs well in image-based tasks like spectrogram classification. It focuses solely on spatial feature extraction, and thus it is an appropriate point of comparison for our CRNN, which incorporates temporal modeling.

As shown in Figure 3, the baseline model ended up performing much better than our CRNN model. It had a validation accuracy of 0.7830 and a test accuracy of 0.7899, which is an 18% increase in accuracy when compared to our CRNN model.

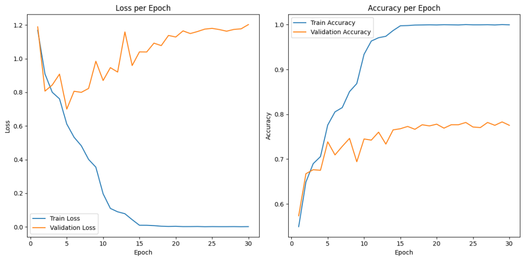

Figure 3: Traing and validation accuracy of ResNet-based CNN baseline model along with loss.

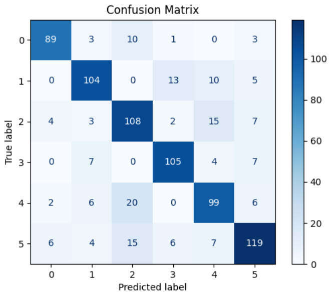

Figure 4: Confusion matrix of ResNet-based CNN baseline mode. The six categories (0-5) represent the emotions neutral (0), happy (1), sad (2), anger (3), fear (4), and disgust (5).

8.2 CRNN Performance

After training our model on three datasets (TESS, CREMA-D, and RAVDESS), our CRNN model achieved a validation accuracy of 0.64% and a test accuracy of 61% as shown in Figure 5.

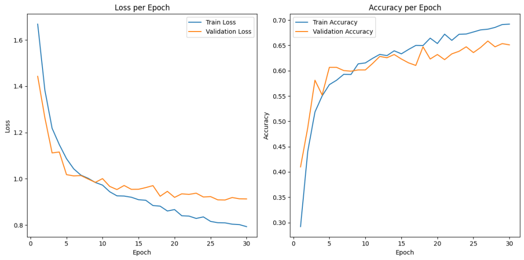

Figure 5: Traing and validation accuracy of CRNN model along with loss.

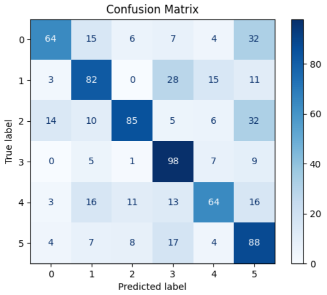

Figure 6: Confusion matrix of CRNN model. The categories are the same as Figure 4.

9 Discussion

Our CRNN model underperformed, achieving 61% test accuracy compared to the baseline ResNet's 79%. This is likely due to the dataset's simplicity, which led the RNN layers to overfit and reduced generalization.

CRNN's strength in modeling spatial and temporal features was underutilized, as fixed spectrogram window sizes likely limited effective temporal capture. Meanwhile, the baseline ResNet, optimized for spatial features, was sufficient for emotion classification without relying on temporal data.

9.1 Performance Along Other Dimensions

Switching to a simpler CNN model improved test accuracy to 75%. This is because CNNs excel at handling static image data, like spectrograms, by focusing on spatial patterns and structures without the complexity of temporal dependencies. Thus, classification performance is improved when the model focuses on the task of image processing.

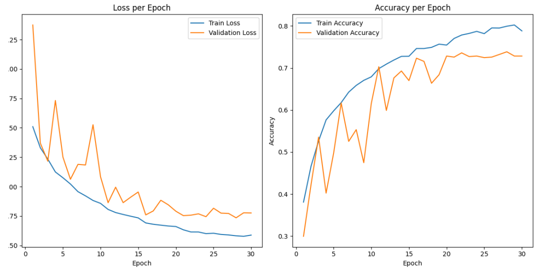

Figure 7: Loss and accuracy of traing and validation data of a CNN model.

9.2 Training Using TESS/CREMA-D but Testing on RAVDESS

To test real-world performance, we trained our CRNN on TESS and CREMA-D and tested on RAVDESS, achieving only 25% accuracy. Most predictions were classified as "angry" due to differing voice patterns in RAVDESS. This suggests our model, with limited training data, wouldn't generalize well in real-world scenarios.

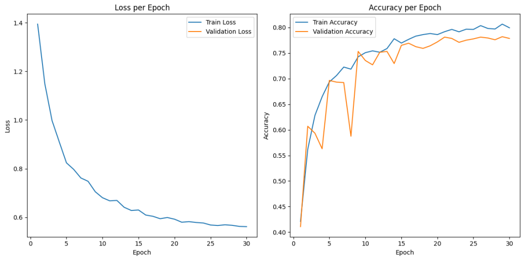

Figure 8: Loss and accuracy of training and validation of CNN model trained on TESS and CREMA-D and tested on RAVDESS.

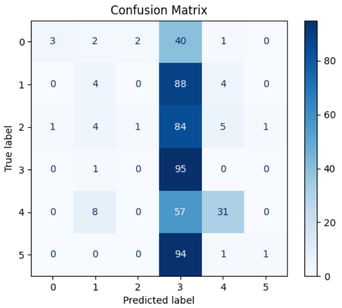

Figure 9: Confusion matrix of CNN model trained on TESS and CREMA-D and tested on RAVDESS. The categories are the same as Figure 4.

9.3 Overfitting With the TESS Dataset

Training and testing our CRNN on just the TESS dataset yielded 99% accuracy, thanks to its balanced yet homogeneous samples from only two speakers. Conversely, when using three datasets (TESS, CREMA-D, RAVDESS), the model struggled to generalize across varied recording conditions, speakers, and emotions, leading to more average performance.

10 Conclusion

For the tasks of SER, the CRNN model has demonstrated a mediocre performance when classifying data across the TESS, RAVDESS, and CREMA-D datasets. Moreover, it heavily struggled with new unseen data. Although the CRNN model achieved overall poor performance compared to a simpler CNN model, it's possible that better results could be obtained with more diverse data. Additionally, if the team had more computing power to train the CRNN model using the audio waves rather than relying on spectrograms, the CRNN model might have performed better, especially given the memory constraints we faced during this project.