SPEECH EMOTION RECOGNITION USING DEEP LEARNING

Hazel Sharma
Student# 1007125200
hazel.sharma@mail.utoronto.ca

Nicole Philip
Student# 1006824288
nicole.philip@mail.utoronto.ca

Kevin Chang
Student# 1008134083
kj.chang@mail.utoronto.ca

Alex Lau
Student# 1007003920
aylex.lau@mail.utoronto.ca

ABSTRACT

The ”Speech Emotion Recognition” project utilizes deep learning techniques to classify human emotions present in audio speech recordings. This is a final report for the project, detailing data processing, architecture, model results, discussion and ethical considerations.

—–Total Pages: 9

1 INTRODUCTION

In recent years, speech recognition systems have become ubiquitous in our everyday lives, with technologies such as virtual assistants and voice activated devices making life more convenient and accessible, particularly for persons with disabilities (Champion, 2023).

These recent advancements have been due to the use of machine learning approaches, which have been applied to speech recognition tasks with great success, but often only focus on the content of the dialogue (Son). However, in the case of human speech, the tone and emotion of the speaker being conveyed can be just as important as the content of their words (Int).

Thus, the goal of this project is to train a deep learning model that can reliably recognize the emotions or feelings of a person speaking. Given an audio recording, the model will select the most suitable matching emotion from a predetermined set of labels.

Many factors, such as the level of variability in the data used to train the model, and aspects such as the overall volume, audio quality, and accents of the recorded speakers, make this a difficult problem to solve using traditional computational methods. Therefore, a deep learning approach which uses a neural network to make predictions was used for this task.

2 ILLUSTRATION/FIGURE

Figure **1** is an illustration representing the final model architecture. It showcases the process of extracting relevant features from an audio sample, inputting it to the CNN model, and the final output produced.

3 BACKGROUND & RELATED WORK

Speech Emotion Recognition (SER) is a widely researched area which has various methods and algorithms developed to detect emotion. Traditional non-neural network methods such as Hidden Markov Models (HMM) and Support Vector Machine (SVM) can be used in tandem to effectively

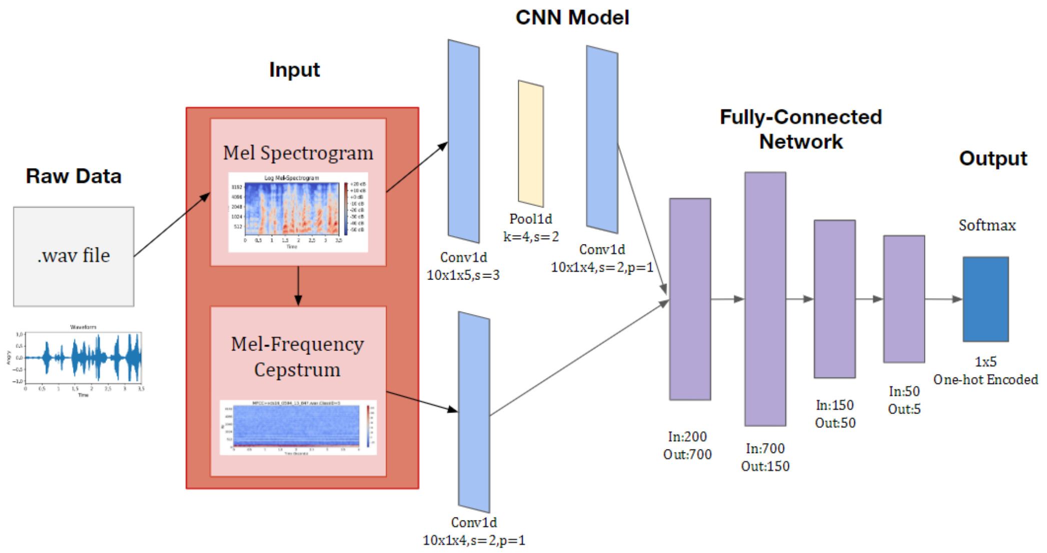


Figure 1: Final Model Architecture

detect emotions in speech and classify them into different categories. This enhances the accuracy of speech emotion recognition systems (Abbaschian et al., 2021).

However, in recent years, the trend has shifted to deep learning-based models like Convolutional Neural Networks (CNN). An example of a renowned CNN architecture is AlexNet. It is one of the existing approaches in the development/research of SER systems and is known to be a pioneer for image classification tasks. It can be adapted to SER systems by treating speech data as RGB spectrogram images, or other time-frequency representations, to differentiate between emotional classes. Then, to classify the emotions, AlexNet is trained on labeled emotional speech data and since it is an already pre-trained network, the process of fine tuning is much faster (Lech et al., 2020).

Using SER systems, there has been significant exploration in the field of Robotics, ranging from children's toys to emotyping wristwatches. Prominent examples include Anki's Cozmo , a domestic toy robot (McStay & Rosner, 2021) , and Softbank's Pepper, a humanoid robot (Do, 2022). They use cameras and microphones to gain information and apply behavior-based architecture with neural networks to identify/learn the user's emotions. Prediction of emotion has also been used to detect high levels of fear, anxiety and anger in hospital patients for psychological purposes as well as in businesses for market research to analyze positive/negative responses to products (Kerkeni et al., 2019). However, companies have faced many challenges when implementing this technology in real-life due to factors such as linguistic diversity, age variations, audio quality, and more.

For our project, we plan to focus on detecting emotion through English speech samples spoken by adults.

4 Data Processing

This section covers the data collection, cleaning and transformation process.

We used four different data sets to train our model. Each data set has been widely used and are reliable, as they were obtained from credible institutions such as the University of Toronto. Combined, they offer a diverse data set with almost 10,000 audio files. A combined data set provides a wide spectrum of speakers of different genders, age, and ethnicity, which reduces bias and aids in transfer learning. Table **1** lists the key characteristics of the data sets.

The following sections detail the steps taken to clean and format the raw audio files into a usable input for the model.

Table 1: Key characteristics of datasets CREMA-D (Lok, 2019a), RAVDESS (Livingstone, 2019), SAVEE (Lok, 2019c), and TESS (Lok, 2019b)

Data Set	**# of Clips**	**# of Speakers Recorded**	**Gender Distribution**	**Age Distribution**	**Ethnicity Distribution**
Crowd Sourced Emotional Multimodal Actors Dataset **(CREMA-D)**	7442	91 actors	48 male 43 female	20 - 74	African American, Asian, Caucasian, Hispanic, and Unspecified
Ryerson Audio-Visual Database of Emotional Speech and Song **(RAVDESS)**	1440	24 actors	12 male 12 female	-	North American Accent
Surrey Audio-Visual Expressed Emotion **(SAVEE)**	480	4 students & researchers	4 male	27 - 31	Native English Speakers
Toronto Emotional Speech Set **(TESS)**	2800	2 actresses	2 female	26, 64	-

4.1 SETTING UP DATA SETS

After identifying the data sets, we combined them into a single data set to create a data frame which consists of the path to the audio file and a label indicating its corresponding emotion. The data frame was then saved as a .csv file. All data sets contained .wav audio files. Every audio file was organized using the syntax in their filename, it contained all the information about the speech sample (ex. Gender of speaker, sentence type, actor number). The name was then parsed to extract the emotion associated with the file. Unfortunately, all data sets used different naming conventions so first, we had to operate on them individually by creating separate data frames for each. After the separate data frames were created, we plotted a graph to display how the emotions were distributed across audio files within each data set as seen in Figure **2**.

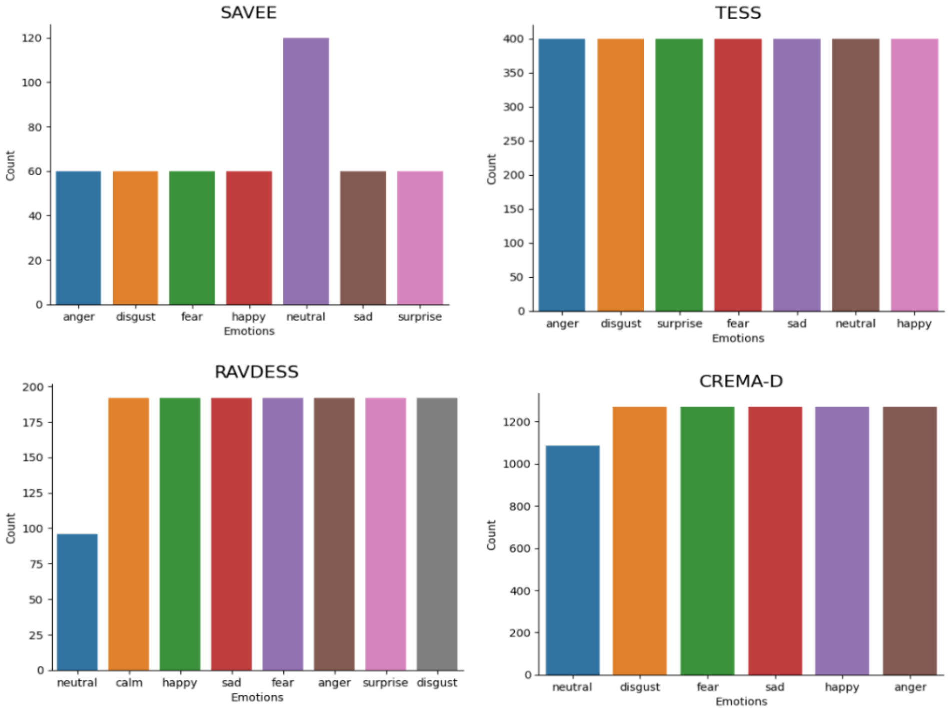


Figure 2: Count of emotions in RAVDESS, SAVEE, TESS and CREMA-D data sets

Using the plots, we were able to identify the emotions that were not common across the data sets as well as the emotions making the data set unbalanced because of its excessive or inadequate number of audio files. For example, the RAVDESS data set has an extra emotion called “Calm” whereas the others did not. This resulted in the ”Calm” emotion having a smaller number of audio files when the data sets are combined. Thus, to ensure the accuracy of the model, only emotions common across all four data sets were used. Similarly, the emotions “Surprise” and “Neutral” were removed because they would make the distribution of emotions unbalanced, thus possibly resulting in a biased set.

Overall, our final processed data set comprised of 9615 audio files with 1923 files per emotion with the following 5 emotion classes: “Happy”, “Sad”, “Anger”, “Fear”, and “Disgust” as seen in Figure **3**.

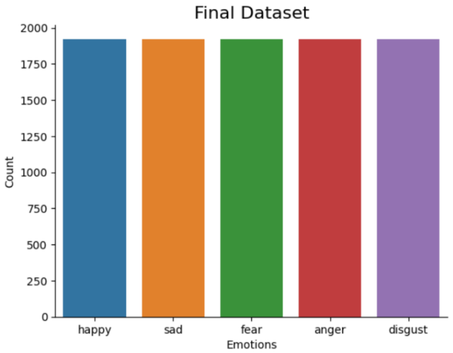


Figure 3: Emotion distribution of the combined data set

After merging the data sets and removing the extra emotions, we tried to access specific rows of the data frame we created, however would get an error. Through debugging, the problem was determined to be the sequencing of row numbers as the index was not reset. Furthermore, when the data sets were combined, a graph was generated to display the new emotion distribution. However, the number of emotions per category did not increase as expected, instead the number of emotion categories increased. We then realized this was because all data sets labeled their emotions differently. For example, “Sad” vs. “Sadness” was considered as two different categories, so the labels were changed to make it consistent, which fixed the problem.

4.2 FEATURE EXTRACTION

After combining and processing the data sets, we proceeded with feature extraction to gather meaningful information from the audio signals. In many academic papers on audio classification, the prevalent features included energy, pitch, MFCC’s, Mel-Spectrograms, Band Energy Ratio, Zero-Crossing Rate and many more Kerkeni et al. (2019).

For our project, the focus was on the frequency-domain features Mel Frequency Cepstrum Coefficients (MFCC) and Mel-Spectrograms (Mel-Spec) as they are highly associated with the human perception of speech (Venkataramanan & Rajamohan, 2019). Librosa and TorchAudio Python libraries were used to load each audio file as a floating point time series and compute the coefficients/spectrogram.

A Mel-Spectrogram is a visual depiction of the spectrum of frequencies. It is calculated using Fast Fourier Transforms (FFTs) and maps the results onto the Mel-frequency scale. Since the human auditory system is good at discerning small changes in pitch at low frequencies compared to high ones, the Mel-scale helps our features match closely to what humans hear (Meng et al., 2019).

MFCCs, on the other hand, are the logarithmic perception of intensity and tone. They convey distinct units of sound (Phemones) and are known to reflect the shape of the vocal tract. It is generated by applying direct cosine transforms on Mel Spectrogram data (Meng et al., 2019). Both features thus hold a lot of useful information that would help the model classify emotions.

Being unfamiliar with the mathematical concepts of audio processing as well as the Librosa commands made it more difficult when analyzing the spectrograms we generated. as we could not tell if the images produced were correct. To ensure our spectrograms were correct, we went through different documentation and compared our spectrograms to example industry standard spectrograms.

Finally, we stored each feature array into a column in the data frame in a row associated with its emotion label in a comprehensive .csv file. Below in Figure **4** and **5**, you can see example visual representations (an Audio Waveform, an MFCC and a Mel-Spectrogram) of a cleaned audio sample saying ”Dogs are sitting by the door” by a male speaker in the emotion ”Happy”.

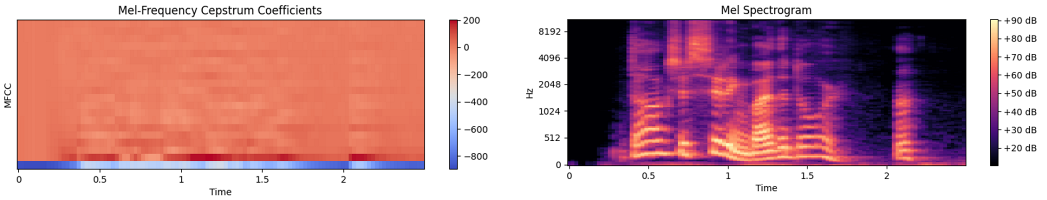


Figure 4: MFCC and Mel-Spectrogram of a cleaned audio sample saying ”Dogs are sitting by the door” by a male speaker in the emotion ”Happy”

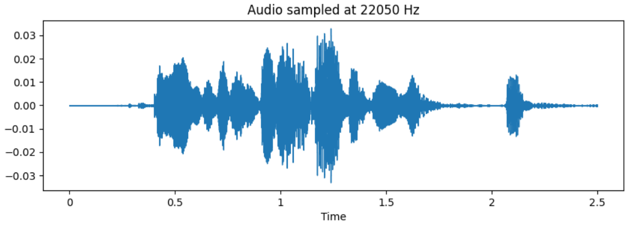


Figure 5: Audio Waveform of a cleaned audio sample saying ”Dogs are sitting by the door” by a male speaker in the emotion ”Happy”

5 ARCHITECTURE

During the term, our team worked on various model architectures for Speech Emotion Recognition. We will focus on the CNN and SVM due to the time spent on those models as well as the interesting results. Please note, a dual Recurrent Neural Network with separate RNNs for Mel-Spectrogram and MFCC data, each consisting of two layers, followed by a fully connected network was also developed by the team.

5.1 FINAL CNN MODEL

Given that each sample consists of a 1x128 vector for Mel Spectrogram, a 1x20 vector for MFCCs, and a truth label, the first step was to convert the truth label to a one-hot encoding. This transformed the truth label to a 1x5 vector, and allowed for comparison to the output of the model through a Cross Entropy Loss function, suitable for multi-class classification. As both vectors in a sample contain different but complimentary information, we used a convolutional neural net to extract features from both and then combine them using a fully-connected network to map the features into a prediction (output). As both vectors are 1D, we used the 1D-Convolution and 1D-Pooling layers from PyTorch. The outputs from the convolutional layers were then concatenated and passed into our fully-connected network.

More specifically, for the 1x128 Mel spectrogram vector, the vector was passed into a 1D convolutional layer with 10 output channels, kernel size of 5, and stride of 3, then into a 1D max pooling layer, with kernel size 4 and stride of 2, and finally into another 1D convolutional layer with 10 output channels, kernel size of 4, stride of 2, and padding of 1. This produced a 10x10 output tensor.

Due to the smaller vector size, the 1x20 vector representing the MFCCs was only passed through one convolutional layer with 10 output channels, kernel size of 4, stride of 2, and padding of 1, also producing a 10x10 output tensor. These two tensors were concatenated into a single 20x10 tensor, and passed into a fully connected network.

The fully connected network consisted of four layers, the first with 700 neurons, the second with 150 neurons, then 50, then 5. This final layer produced the output prediction, and the label was then compared to the output.

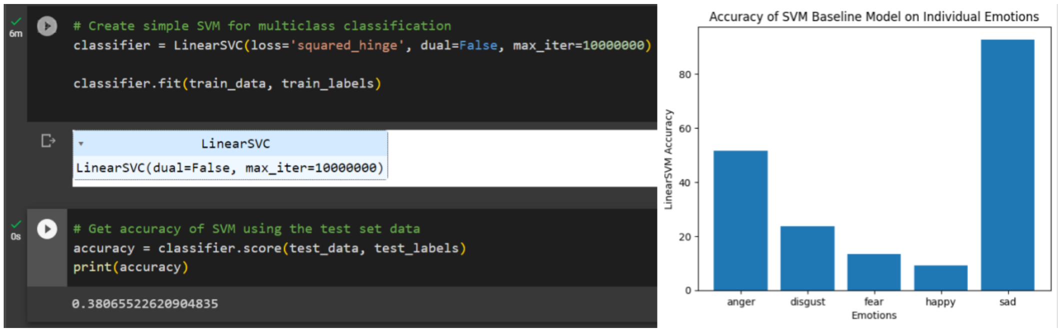
5.2 Baseline Model

The baseline model chosen for this project is a simple machine learning model based on the Support Vector Machine (SVM) algorithm, using the Scikit-Learn library.

Machine learning models based on this architecture have shown to be effective in audio classification problems, including for non-binary classification tasks, and are generally much more effective than non-machine learning approaches to audio classification (Chen et al., 2006).

The spectrogram data and its truth labels were retrieved from a .csv file, after which it was split into testing and training segments, and the labels were normalized. The model then iterates and fits the data of the training set (in numpy array format) to their corresponding labels. Once the SVM model finished training and converged, it was then tested using the test dataset. Using this process, the baseline model achieved an accuracy of 38% on the test dataset with 5 truth labels, which is superior to a model which performs randomly (1/5 chance = 20.0%).

Figure **6** displays the code segment used for model fitting and testing as well as the baseline model's accuracy on individual emotions.



```
# Create simple SVM for multiclass classification
classifier = LinearSVC(loss='squared_hinge', dual=False, max_iter=10000000)

classifier.fit(train_data, train_labels)
```

```
LinearSVC
LinearSVC(dual=False, max_iter=10000000)
```

```
# Get accuracy of SVM using the test set data
accuracy = classifier.score(test_data, test_labels)
print(accuracy)
```

```
0.38065522620904835
```

Figure 6: Screenshot of LinearSVM model fitting/testing, and graph of the baseline model's accuracy on individual emotions.

6 Results & Discussion

6.1 Quantitative Results

Several quantitative measures were chosen to evaluate our model.

6.1.1 Training and Validation Accuracy

Across the 3 different models that were created, the Convolutional Neural Network achieved the highest training and validation accuracies of 85% and 62% respectively. In Figure **7**, the graph shows the CNN's model training accuracy consistently improved over epochs, which indicates that the model effectively learns to extract spatial features from the Mel-Spectrogram and MFCC data. However, the validation curve begins to plateau at early epochs indicating a possible chance of overfitting. Our network may have memorized the training input, making it harder to generalize on unseen data.

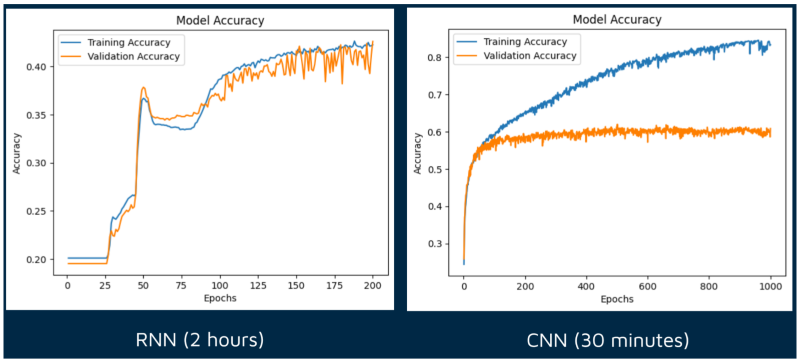


Figure 7: Graphs of the RNN and CNN Model Accuracy

In contrast, the RNN model displays a different pattern and reaches a training accuracy of approximately 45%. Unlike the CNN, validation accuracy maintains a closer proximity in performance to the training accuracy, indicating better generalization. This suggests that the RNN's temporal memory aids in grasping the evolving emotional cues present in the speech data.

6.1.2 Precision, Recall & Confusion Matrix

A confusion matrix, as shown in the Figure **8**, was created for our test data set to evaluate the performance for each class. When comparing the model's predictions to the true labels, the emotions Anger and Sadness achieved the highest accuracies of 82% and 85% respectively, even though the data set was balanced. We noticed our model is able to effectively identify and classify Anger and Sadness within our testing data set compared to other emotions. Those emotion classes also have the highest precision and recall values.

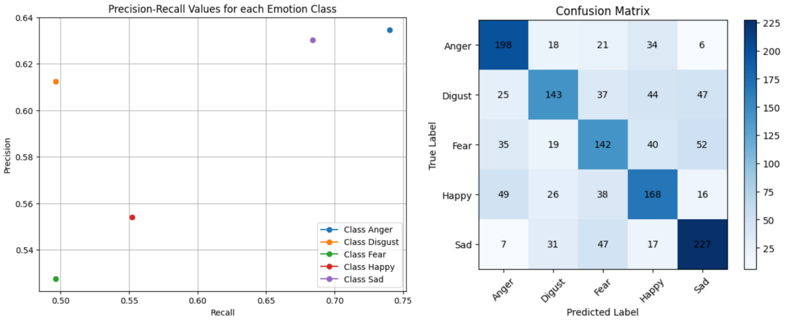


Figure 8: Graphs of the Precision/Recall values for each emotion class and a Confusion Matrix of the testing results

6.2 Qualitative Results

Figure **9** showcases the two spectrogram samples processed in our Google Collab for a Sadness and Anger audio sample.

In the qualitative visualizations of the spectrogram, we can see certain emotional characteristics represented. While the emotion of anger is often linked to high-frequency, loud and intense speech, sadness is known to involve slow dynamics and low pitches. Thus, in the spectrogram scales we find that the angry audio has a sound intensity level 30db higher than the sadness audio.

Anger and Sadness, two opposite emotions have distinct auditory cues that could have enabled the model to discern them with greater accuracy. This is further supported by our confusion matrix as Anger and Sad samples were uncommonly mistaken for the other emotion and have the smallest amount of false positive/negatives ($\leq$10 across 4000 samples).

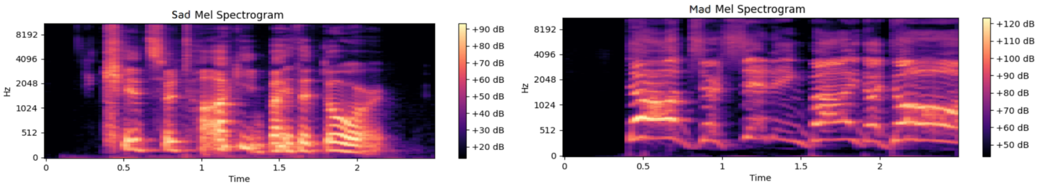


Figure 9: Mel-Spectrograms of a Sadness and Anger audio sample saying the phrase ”My shoulders felt as if they were broken”

Conversely, classes with relatively lower accuracy correspond to emotions that encompass intricate or contextually nuanced cues that are not clearly indicated by the spectrogram. Audio samples with emotions like ’Surprise’ or ’Disgust’ might both involve similar rapid shifts in pitch and energy, which could challenge the model’s ability to capture subtle temporal variations.

6.3 Evaluate Model on New Data

In order to test our model on new data, we split our combined data set into 70% for training, 15% for validation and 15% for testing. Each set had unique samples, spoken by a variety of speakers of different ages, genders and spoken phrases. This allowed no two audio samples to be exactly the same. In addition, we also tested our model on a completely new and unseen emotional speech database in New Zealand English by JL Corpus (JL-Corpus).

Table 2: Various testing accuracies across our Models and comparison to Human Emotion Recognition Accuracy (Tsai & Chen, 2022)

Model	SVM Baseline	RNN	CNN	Human Ear
Testing Accuracy	38%	43%	62%	72%

As only validation data was used to tune the hyperparameters, the test performance was only taken into account at the end of the training process. Over the term, we worked to improve our training and validation accuracy over various models and our final CNN model achieved the highest testing accuracy of 62%. The model was able to yield a test accuracy that was close to the Human Emotion Recognition accuracy (Tsai & Chen, 2022).

6.4 Discussion

Our final, most refined model using a CNN architecture was able to reach near-human level accuracy in determining speaker emotions while greatly outperforming the baseline model. CNNs, being prominently used in image processing, turned out to be the best architecture to use as our problem of classifying audio ultimately became an image/pattern recognition task.

Human speech is multifaceted and analyzing separate aspects such as the volume, words spoken, and length of the clip on their own is simply not enough to form a solid conclusion. By converting audio samples into Mel-Spectrograms, we believe that we have found a more holistic approach to detecting the nuanced differences in emotions which takes all factors into account at once. Specifically, our CNN model excelled in cases where emotions were characterized by distinct spectrogram patterns. The MFCCs and Mel-spectrograms were able to capture sharp emotional differences, causing the model to better recognize specific emotions. This explains the high accuracies for Anger and Sadness audio samples. To further improve the model, it would be better to extract other features from the audio sample such as Zero Crossing Rate, Spectral Centroid and Spectral Rolloff to help it identify more characteristics relevant for the other 3 emotions.

When improving our initial CNN model, we made changes such as increasing the dataset size, adding additional layers to the fully connected network, and reducing the number of emotion classes to balance the dataset and reduce complexity. These adjustments significantly raised our model's performance.

However, despite achieving near-human accuracy and using a suitable model architecture, a testing accuracy of 62% still leaves much room for improvement. As the model was trained from scratch using our own set of model parameters, we believe that given more time, an even better model could be trained if we had used a set of pretrained weights from previous audio classification datasets to initialize our neural network. In addition, compared to the CNN, the RNN took much longer to run, requiring around 2 hours of run time for 200 epochs, while the final CNN only took 30 minutes for 1000 epochs. If we had more time, it would be interesting to see if the RNN can exceed the CNN's accuracy if given a larger data set to train with as well as testing different combinations of hyperparameters.

7 Ethical Considerations

Though this project was conceived without malicious intent, there are still ethical considerations that should be taken into account.

7.1 USAGE

Firstly, the model may be used in ways which will have ethical ramifications. For example, the model could be used in conjunction with law enforcement in order to detect arguments (categorized by an angry emotional reaction). If our model provides information used in any capacity by law enforcement or the government, this will give rise to concerns such as whether utilizing these devices for surveillance is authoritarian. In addition, if this information is used in the context of a police investigation, we may be partly responsible in directing the course of the investigation.

7.2 MISAPPROPRIATION

Another consideration is that our model may be altered in order to serve purposes that it was not intended for. As an example, identifying audio features can ultimately be extended to identifying distinct words. The privacy violations made possible by such an innovation would definitely be worthy of concern.

7.3 ACCURACY

As a machine learning model, the precise accuracy on truly new data is difficult to ascertain. Before applying our model to real life situations, and making guarantees about the accuracy rate, it would need to be trained and tested on a larger data set than what was used during the course of the project.

8 Google Colab Link

A Google Colab notebook will be used for project collaboration. See: `https://colab.research.google.com/drive/1os61di-X0zk1Nh9El65hyYUd08K3Y8VC?usp=sharing`

References

URL `https://doctorsspeakup.com/content/intonation#:~:text=Intonation%20is%20very%20important%20in,a%20statement%20and%20a%20question).`

URL `https://sonix.ai/history-of-speech-recognition`.

Babak Joze Abbaschian, Daniel Sierra-Sosa, and Adel Elmaghraby. Deep learning techniques for speech emotion recognition, from databases to models. *Sensors*, 21(4):1249, 2021. doi: 10.3390/s21041249. URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7916477/`.

Vicky Champion. How does voice recognition help people with disabilities?, Mar 2023. URL https://getaboutmobility.com.au/blogs/news/voice-recognition-help-people-with-disabilities#:~:text=Using%20Computers%3A%20Voice%20recognition%20is,notes%20on%20computers%20and%20laptops.

Lei Chen, Sule Gunduz, and M. Tamer Ozsu. Mixed type audio classification with support vector machine. In *2006 IEEE International Conference on Multimedia and Expo*, pp. 781–784, 2006. doi: 10.1109/ICME.2006.262954.

Liz Do. Meet pepper: An ai robot that will reduce wait times in hospitals, Feb 2022. URL https://news.engineering.utoronto.ca/meet-pepper-an-ai-robot-that-will-reduce-wait-times-in-hospitals/.

Hazalkl JL-Corpus. Hazalkl/jl-corpus repository of emotion speech audio files. *DagsHub*. URL https://dagshub.com/hazalkl/JL-Corpus.

Leila Kerkeni, Youssef Serrestou, Mohamed Mbarki, Kosai Raoof, Mohamed Ali Mahjoub, and Catherine Cleder. Automatic speech emotion recognition using machine learning. In Alberto Cano (ed.), *Social Media and Machine Learning*, chapter 2. IntechOpen, Rijeka, 2019. doi: 10.5772/intechopen.84856. URL https://doi.org/10.5772/intechopen.84856.

Margaret Lech, Melissa Stolar, Christopher Best, and Robert Bolia. Real-time speech emotion recognition using a pre-trained image classification network: Effects of bandwidth reduction and companding. *Frontiers in Computer Science*, 2, 2020. ISSN 2624-9898. doi: 10.3389/fcomp.2020.00014. URL https://www.frontiersin.org/articles/10.3389/fcomp.2020.00014.

Steven R. Livingstone. Ravdess emotional speech audio, Jan 2019. URL https://www.kaggle.com/datasets/uwrfkaggler/ravdess-emotional-speech-audio.

Eu Jin Lok. Crema-d, Aug 2019a. URL https://www.kaggle.com/datasets/ejlok1/cremad.

Eu Jin Lok. Toronto emotional speech set (tess), Aug 2019b. URL https://www.kaggle.com/datasets/ejlok1/toronto-emotional-speech-set-tess.

Eu Jin Lok. Surrey audio-visual expressed emotion (savee), Sep 2019c. URL https://www.kaggle.com/datasets/ejlok1/surrey-audiovisual-expressed-emotion-savee.

Andrew McStay and Gilad Rosner. Emotional artificial intelligence in children's toys and devices: Ethics, governance and practical remedies. *Big Data & Society*, 8(1): 2053951721994877, 2021. doi: 10.1177/2053951721994877. URL https://doi.org/10.1177/2053951721994877.

Hao Meng, Tianhao Yan, Fei Yuan, and Hongwei Wei. Speech emotion recognition from 3d log-mel spectrograms with deep learning network. *IEEE Access*, 7:125868–125881, 2019. URL https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8817913.

Ming-Fong Tsai and Chiung-Hung Chen. Enhancing the accuracy of a human emotion recognition method using spatial temporal graph convolutional networks - multimedia tools and applications, Aug 2022. URL https://link.springer.com/article/10.1007/s11042-022-13653-x.

Kannan Venkataramanan and Haresh Rengaraj Rajamohan. Emotion recognition from speech, 2019. URL https://arxiv.org/abs/1912.10458.