Vocal Emotion Analysis

Capstone III - Final Report

Springboard
Data Science Career Track

Tara Crutchfield
February 25, 2021

Table of Contents

Appendix

1. Objective

Voice recognition technology is all around us; virtual assistants like Apple's Siri and Amazon's Alexa have made their way into every house and pocket, while automated call responders have become the voice behind almost every large company's telephone line. Due to its many applications and increasing usage, voice recognition technology is constantly improving, both in terms of identifying and understanding human speech. As such, one of many avenues to help enhance this technology, is the ability to perceive emotions within a person's voice.

Creatures of all kinds are able to pick up emotional cues through both a combination of inherent inborn skills and years of training, often through childhood development. As such, emotions are an inherently biological characteristic and tend to appear different between cultures and even species. It's also worth noting that emotions are often recognized by a combination of voice qualities, facial features, and body posture.

These complexities pose great challenges to an artificial intelligence's ability to potentially perceive emotions, especially when isolating voices from the face of the speaker. Like any challenge, though, the best way to move forwards is to start small. As such, the purpose of this project was to train a convolutional neural network to recognize emotions from spoken phrases, focusing on North American english.

2. Data

The data used in this project came from two different voice databases: the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) by Ryerson University in Toronto, Canada, and the Crowd-sourced Emotional Multimodal Actors Dataset (CREMA-D) by the Cheyney University of Pennsylvania.

CREMA-D was composed of 7442 recordings by a total of 91 actors. Each actor vocalized one of 12 phrases using 6 different emotions: **Neutral**, **Happy**, **Sad**, **Angry**, **Fear**, and **Disgust**. RAVDESS was composed of a total of 7356 files, with 24 professional actors vocalizing 2 sentences using the same emotion as CREMA-D as well as two additional emotions: **Calm** and **Surprise**. Vocalizations were provided in both spoken and song form, and clips were separated as audio only, video only, and audio-video files. For this project only the spoken audio clips were used, giving a total of 1440 clips. Table 1 provides a short breakdown of these collected audio clips.

Table 1: Breakdown of the data from the two databases.

Database	Actors	Phrases	Emotions	Files
RADVESS	24	2	8	1440
CREMA-D	91	12	6	7442
Total	115	14	8	8882

While both databases were composed of acted phrases rather than instances of real emotion, each had a system of rating the clips such that each clip felt both accurate to the emotion and genuine in its delivery. As mentioned before, these databases were limited to North American english so the results of this project may be limited to this geological location.

3. Packages

This project utilized Python (3.8.1) and Jupyter Notebook (7.20.0) for all processes. Pandas (1.2.1) and NumPy (1.19.5) were used for general data extraction and manipulation, and both Matplotlib (3.3.4) and Seaborn (0.11.1) were used for data visualization. The voice activity detection utilized the packages collections, contextlib, sys, wave, and webrtcvad (2.0.10) while the remaining audio processing and feature extraction process utilized librosa (0.8.0) as well as sounddevice (0.4.1) and soundfile (0.10.3). All machine learning was done using Keras (2.4.3) with preprocessing and model metric evaluation utilizing Scikit Learn (0.24.1).

4. Data Wrangling

Both databases provided the audio as WAV files, with each clip's information specified in the filename. Examples of these filenames, as well as a breakdown of their interpretation, can be found in Figure I within the appendix. CREMA-D also provided a separate file, saved as `CREMA_D Actors.csv`, containing extra information on their actors, such as age, race, and ethnicity.

The file names were collected and compiled into two separate datasets, one for each database. These datasets were then cleaned and renamed such that both had the same columns and labeling conventions. These datasets were joined and saved as the file `Audio Legend Raw.csv`, a description of which can be found in Table 2.

Table 2: Breakdown of the file `Audio Legend Raw.csv`

No.	Column Name	Description	Non-Null	Type
0	path	Path to respective database audio	8882	object
1	original	Filename	8882	object
2	actorid	Actor ID	8882	int
3	statement	Voiced statement	8882	int
4	emotion	Clip intended emotion	8882	object
5	sex	Actor sex	8882	object
6	race	Actor race	7442	object
7	ethnicity	Actor ethnicity	7442	object

5. Data Cleaning

5.1. Voice Activity Detection

Voice activity detection VAD was used to eliminate any background noise and isolate the voices in each clip. The VAD has three levels of intensity, and each clip was processed first using the mid-level setting. If a voice was too quiet to be detected by the VAD, the clip was then processed again using the lowest level intensity. If the voice still could not be distinguished from the background noise, this clip was dropped. By the end of the process, 40 clips in total were dropped from the project, a great majority of them falling under the **Sad** label and spoken by **Female Actresses.**

5.2. Extra Label

The second task was to determine how to deal with the two additional labels found in the RAVDESS database: **Calm** and **Surprise.** As RAVDESS was significantly smaller than CREMA-D, these labels were greatly underrepresented. After manually comparing some of the **Calm** and **Surprise** labeled clips with those of the other emotions, I came to the conclusion that **Calm** could easily fit the **Neutral** category but that **Surprise** felt too varied to fit in any of the other labels. Due to the limited size of the category, all 191 **Surprise** clips were dropped**.**

Figure 1 shows the distribution of the labels before and after this change. Overall, combining **Calm** and **Neutral** nicely balanced the dataset, as it helped boost the size of the **Neutral** label to better match that of the other emotions.

*Figure 1: Two bar graphs showing the distribution of the labels before and after changing the label **Calm** to **Neutral** and dropping the label **Surprise**.*

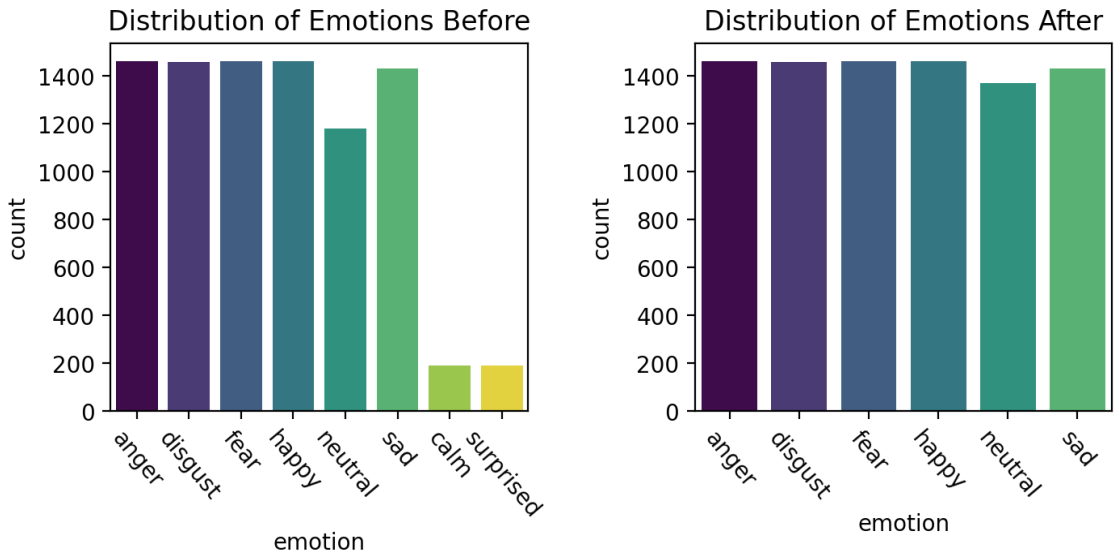


Table 3: Description of the final legend.

No.	Column Name	Description	Non-Null	Type
0	original	Original filename	8651	object
1	actorid	Actor ID	8651	int
2	statement	Voiced statement	8651	int
3	emotion	Clip intended emotion	8651	object
4	sex	Actor sex	8651	object
5	race	Actor race	7307	object
6	ethnicity	Actor ethnicity	7407	object
7	time	Length of clip in seconds	8651	float
8	filename	New filename	8651	object

After both processes, a total of 231 clips were dropped, around 3% of the total amount of clips. The remaining 8651 audio clips were saved under new names for consistency and a new, updated legend was created to reflect the changes made. A brief description of this new legend, saved as `Audio Legend Clean.csv`, can be seen in Table 3.

6. Analysis

A brief analysis was performed on the legend to ensure that the data was well balanced. These factors included the background of the voice actors as well as the distribution of phrases throughout the dataset.

6.1. Phrases

Between both databases, there were a total of 14 phrases used to express emotions. Each phrase was a single sentence that was semantically neutral such that each could be easily acted in any of the eight original emotions. Table 4 presents all of these phrases as well as the integer labels used to denote them in the datasets. In this table, the two phrases above the dashed line (**Phrase 1** and **Phrase 2**) come from RAVDESS, while those below came from CREMA-D.

Figure 2 shows a bar graph of the overall phrase distribution. While many of the phrases appear to have an equal distribution, **Phrase 4** is used twice as often as the others. While at first this was assumed to have been a mistake, deeper investigation concluded that this was indeed the true amount. My best guess as to why this phrase was so common was because the CREMA-D actors got to choose their own lines, and **Phrase 4** was one of the shortest phrases offered. The only issue with this logic is that it would imply that **Phrase 13** should be equally, if not more, common as it is the shortest phrase syllabically.

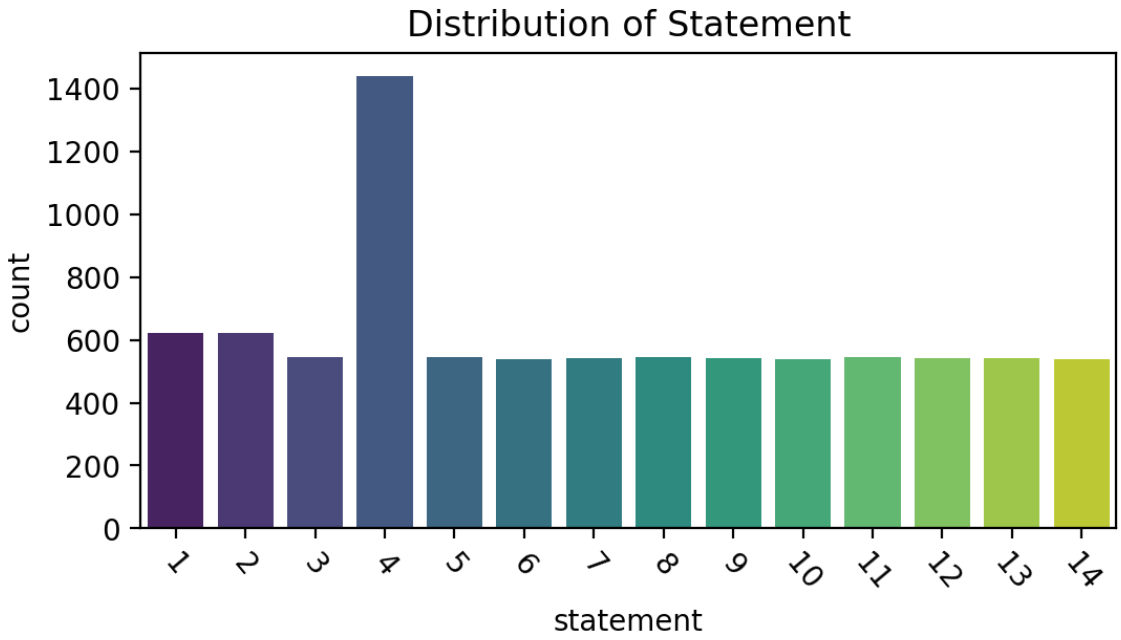
Table 4: Table presenting each phrase and its numeric label. The phrases above the dashed line were used in RAVDESS, while those below were from CREMA-D.

No.	Phrase	No.	Phrase
1	Kids are talking by the door	2	Dogs are sitting by the door
3	Don't forget a jacket	4	It's eleven o'clock
5	I'm on my way to the meeting	6	I think I have a doctor's appointment
7	I think I've seen this before	8	I would like a new alarm clock
9	I wonder what this is about	10	Maybe tomorrow it will be cold
11	The airplane is almost full	12	That is exactly what happened
13	The surface is slick	14	We'll stop in a couple of minutes

Figure 2 shows a bar graph of the overall phrase distribution. While many of the phrases appear to have an equal distribution, **Phrase 4** is used twice as often as the others. While at first this was assumed to have been a mistake, deeper investigation concluded that this was indeed the true amount. My best guess as to why this phrase was so common was because the CREMA-D actors got to choose their own lines, and **Phrase 4** was one of the shortest phrases offered. The only issue with this logic is that it would imply that **Phrase 13** should be equally, if not more, common as it is the shortest phrase syllabically.

While there was some consideration to limit the size of **Phrase 4**, it was ultimately decided that, as long as the distribution of the emotions within the phrase was well balanced, it should not be of too much concern.

Figure 2: Distribution of the 14 statements used throughout the clips.



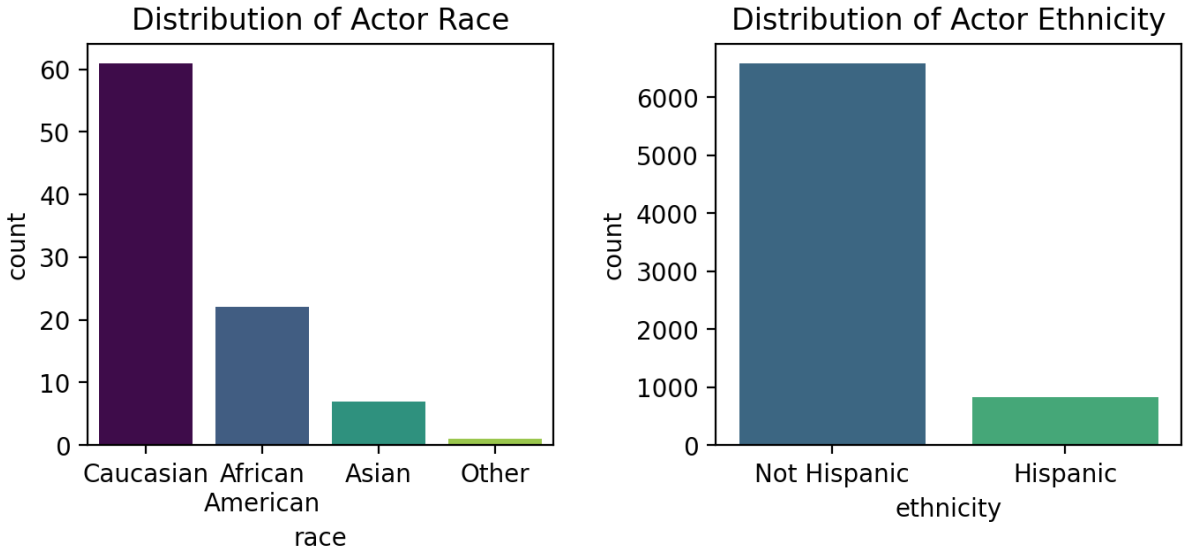
6.2. Race and Ethnicity

Figure 3a shows the distribution of the self-identified race and ethnicity of the actors featured in the dataset. This information was only provided by CREMA-D, so this graph only includes 91 of 115 actors (79%) and only reflects the portion of the data based in the United States.

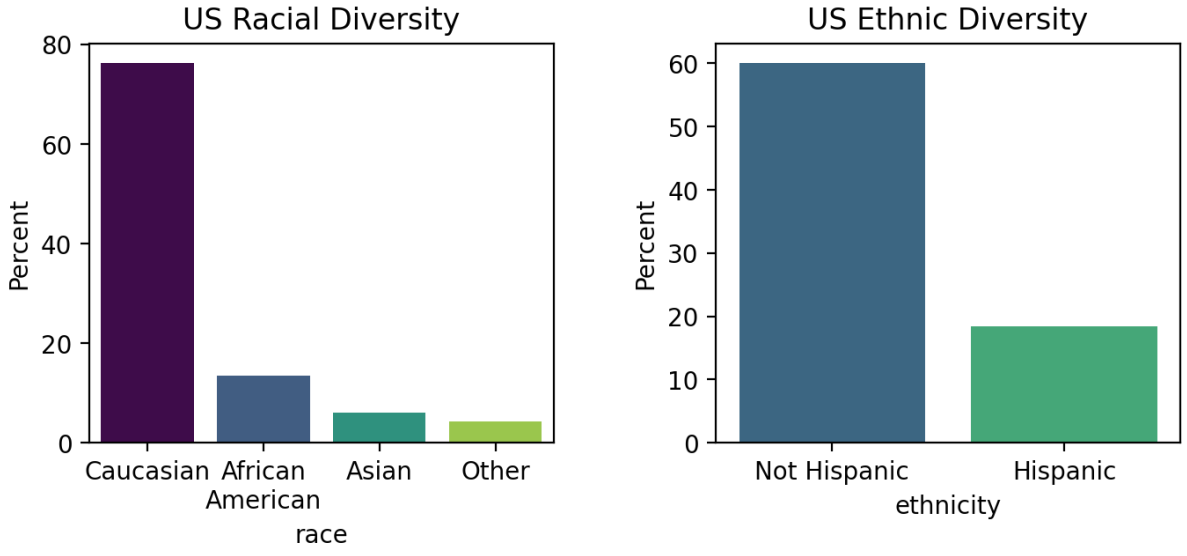
While at first the distribution seems unbalanced, after consulting the latest 2019 U.S. Census, it does appear that the distribution is at least somewhat reflective of the diversity of the USA. For comparison, the diversity of the US population according to this census can be seen in Figure 3b.

Figure 3: Bar graphs showing the distribution of race and ethnicity

a) Within the dataset (CREMA-D)



b) Within the United States

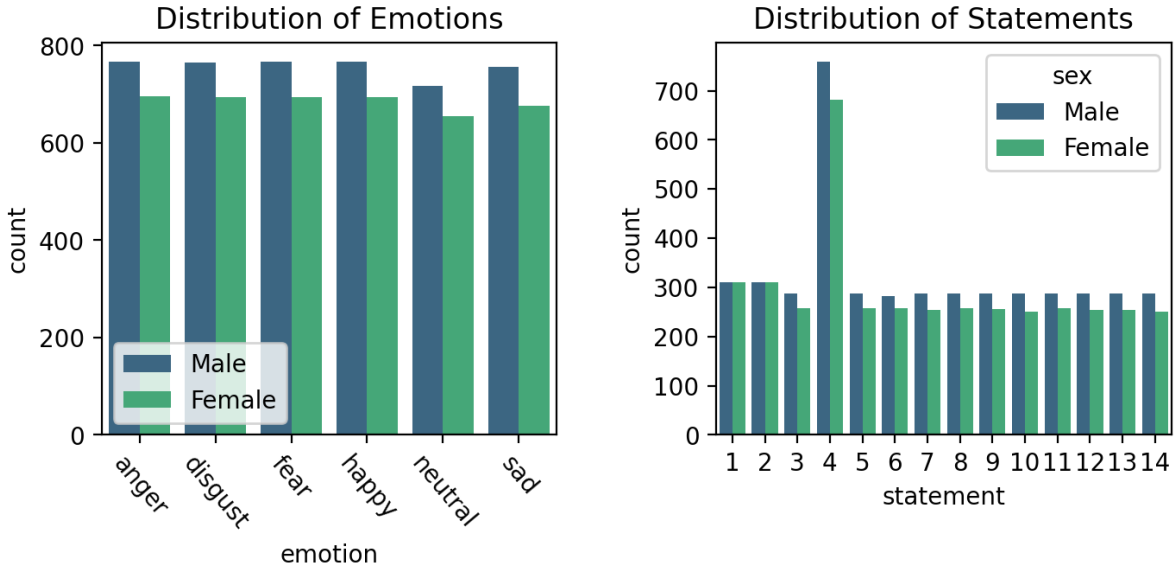


6.3. Gender

Overall, the gender of the actors were quite evenly split, with only five additional men to women. It is worth noting though, that the VAD struggled to pick up the voices of women more often than it did the voices of men, and that this caused more female voiced clips to be dropped during the filtering process.

Figure 4 shows two bar graphs displaying the distribution of gender in both the emotions and statements. While there were definitely more clips featuring male voices, the distributions did not appear to be unbalanced enough to warrant any extra action.

Figure 4: Two bar graphs showing the distribution of gender in both the emotions and statements.

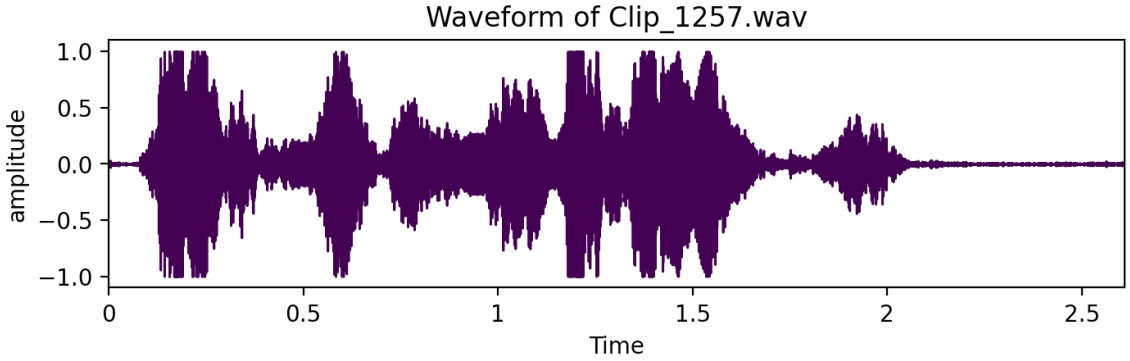


7. Feature Selection

The feature selection process included the extraction of three particular audio qualities from each clip: the mel-frequency spectrogram and cepstral coefficients, as well as the chroma features. These features were saved as two-dimensional images, each with time as it's x-axis.

The following subsections will provide more information on the individual features, as well as provide visualizations utilizing the audio from Clip 1257. This clip came from the CREMA-D databases and includes a **Male Actor** performing **Phrase 8** with **Anger** as the target emotion. Figure 5 shows the waveform of Clip 1257. While this feature was not saved for analysis, it is a good visualization of the audio clip and is provided for comparison to the other features.

Figure 5: Example waveform from Clip 1257



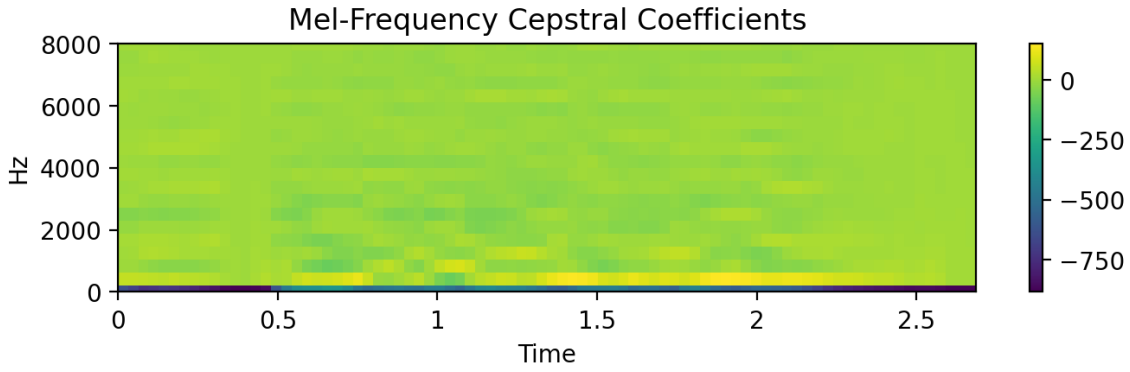
7.1. Mel-Frequency Spectrogram

The mel-frequency spectrogram is a visualization of the frequency spectrum, created utilizing the mel scale, a non-linear pitch scale based on human hearing. Overall, this feature provides an image of how the energy of a signal, or 'loudness', changes over time, focusing particularly on the human perception of these characteristics. An example of the mel-frequency spectrogram can be viewed in Figure II, located in the appendix due to its large size.

7.2. Mel-Frequency Cepstral Coefficients

The mel-frequency cepstrum provides an image of the power spectrum of audio using the previously mentioned mel-scale. Its coefficients can define the overall envelope of the power spectrum, which allows for a visualization of sound based on the shape of the vocal tract. Typically this feature is used to model and understand the qualities of human voices. Figure 6 shows an image of the mel-frequency cepstral coefficients MFCC of Clip 1257.

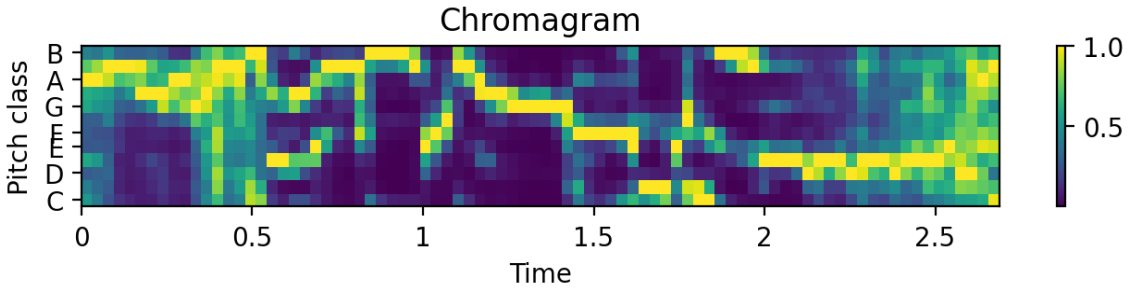
Figure 6: Example MFCC



7.3. Chroma Features

The chroma-based features are created by breaking down the energy of an audio signal into the twelve pitch classes, providing an overall profile of how the pitch shifts with time. For example, the chromagram featured in Figure 7 shows that the voice in Clip 1257 initially wavers at a high pitch before slowly declining to a much lower pitch.

Figure 7: Example chromagram



7.4. Final Model Data

In creating the final dataset, each feature was normalized feature-wise, then stacked together to create a singular image for each entry. Padding was then added such that the final images all had a 160x140 pixels shape, roughly corresponding to an audio clip 4.8 seconds long. Figure III in the appendix gives an example of how this final image appears.

8. Modeling

Two different architectures were tested on two training sets, creating a total of four separate models. Each model started with a max of 50 epochs but was programmed to cease training when the validation loss began to plateau.

The two architectures were similar in that they both shared the pattern of a repeating 2D convolution layer followed by a pooling layer. The style of the pooling layers marked the main difference between the two architectures, with one using maximum pooling MAX and the other using average pooling AVG. Figure 8 provides a flowchart outlining this general architecture.

Figure 8: Flowchart of model architecture.

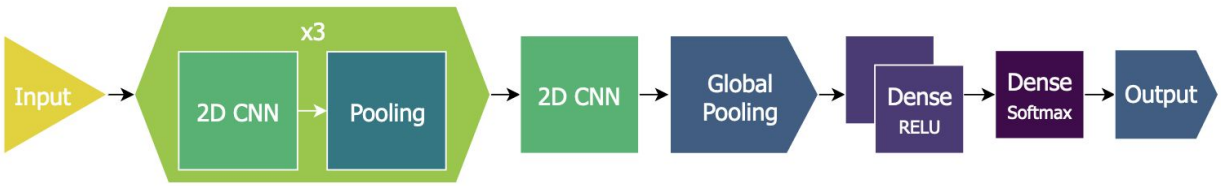


Table 5: Distribution of the training and testing sets

a) Train Set

Gender	Happy	Neutral	Sad	Anger	Disgust	Fear	Total
Female	558	529	541	548	550	555	3281
Male	627	574	604	630	606	598	3639
Total	1185	1103	1145	1178	1156	1153	6920

b) Test Set

Gender	Happy	Neutral	Sad	Anger	Disgust	Fear	Total
Female	137	126	136	148	144	139	830
Male	140	143	152	137	160	169	901
Total	277	269	288	285	304	308	1731

Both architectures were trained twice, with two different training sets. The first of these sets included only the 6 emotion labels EM, while the second combined emotion with the actor gender for a total of 12 separate labels SG. During the initial model analysis, these sets were both split such that 80% of the entries were used for training, while the remaining 20% were used to test model accuracy.

8.1. Emotion Labels

Of the models trained on the six emotion labels, the average pooling model AVG_EM had an accuracy of 57%, while the max pooling model MAX_EM had an accuracy of 56%. Figure 9 shows the results of both of these models with each row annotated such that it displays what percentage of the true label was classified as each emotion. For example, AVG_EM classified 64% of all **Anger** labels correctly but 3% were labeled **Fear**, 14% were labeled **Disgust**, etc.

Overall, AVG_EM had higher average precision, recall, and f1 scores, but when comparing the independent emotions, MAX_EM was slightly better at determining **Disgust**. Both models had a tendency to overlabel emotions as **Sad** and **Happy** and were worst at accurately classifying **Fear.** On the other hand, both models were best at classifying **Anger,** followed by **Neutral.**

Figure 9: Results of the models trained on the emotion labels alone.

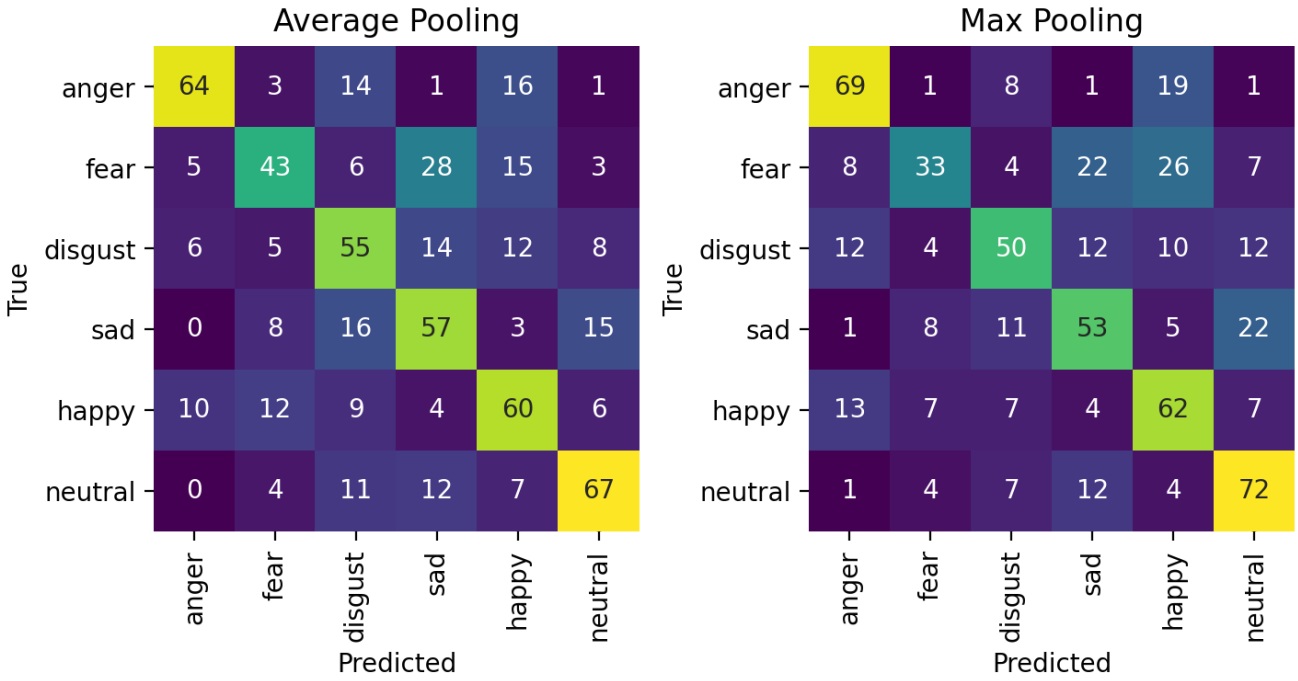
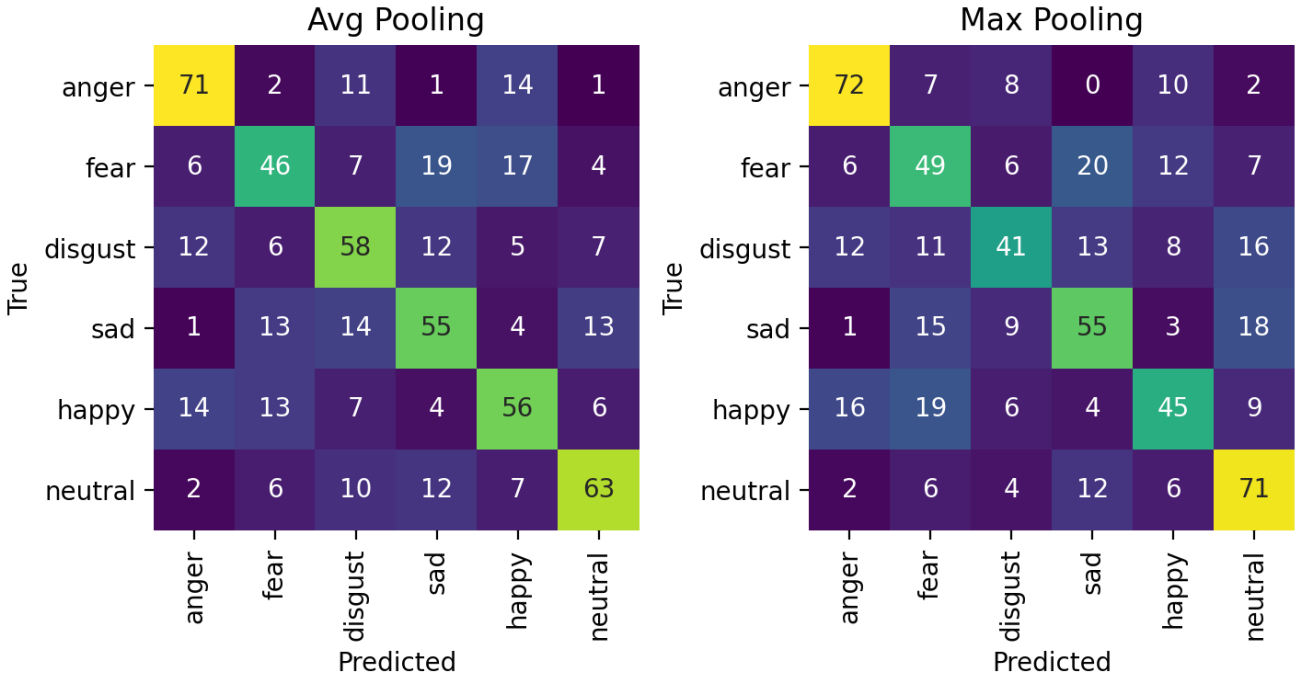


Figure 10: Results of the models trained using both emotion and actor gender.



8.2. Emotion & Gender Labels

Of the models trained on the twelve emotion and gender labels, the max pooling model MAX_SG had an accuracy of 55%, while the average pooling model AVG_SG has an accuracy of 58%. Figure 10 shows a confusion matrix of the results of the emotion labels only, while the full matrix can be viewed in the appendix in Figure IV.

By far AVG_SG was the better of these two models. Despite this, it still had a tendency to overlabel entries as **Sad** and **Happy** and struggled to accurately label **Fear**. Similarly, MAX_SG overelabled entries as **Fear,** and struggled to accurately classify **Happy** and **Disgust.** Like the earlier models, both were also best at determining **Anger,** followed by **Neutral.**

8.3. Results

Table 6 and 7 are a compilation of the overall scores of each of the four models. The first displays the accuracy and average recall, precision, and F1 scores of each. By accuracy alone, AVG_SG is the best of these models, yet it shares the same recall, precision, and F1 scores as AVG_EM, showing that both perform equally well. The second table displays the independent F1 scores of each emotion. Once again, the average pooling models share the highest scores with AVG_SG having the best scores for **Fear, Disgust** and **Sad,** and AVG_EM having the best scores for **Anger, Happy,** and **Neutral.**

Since the two models were so similar in accuracy and tied for the best emotional coverage, it was decided that the final model would be a combination of these two.

Table 6: Scores of all four models as well as the combination.

Model	Accuracy	Recall	Precision	F1 Score
AVG_EM	57%	0.58	0.58	0.58
MAX_EM	56%	0.57	0.57	0.56
AVG_SG	58%	0.58	0.58	0.58
MAX_SG	55%	0.55	0.55	0.55

Table 7: Each model's independent emotion F1 score. The highest scores are highlighted in green.

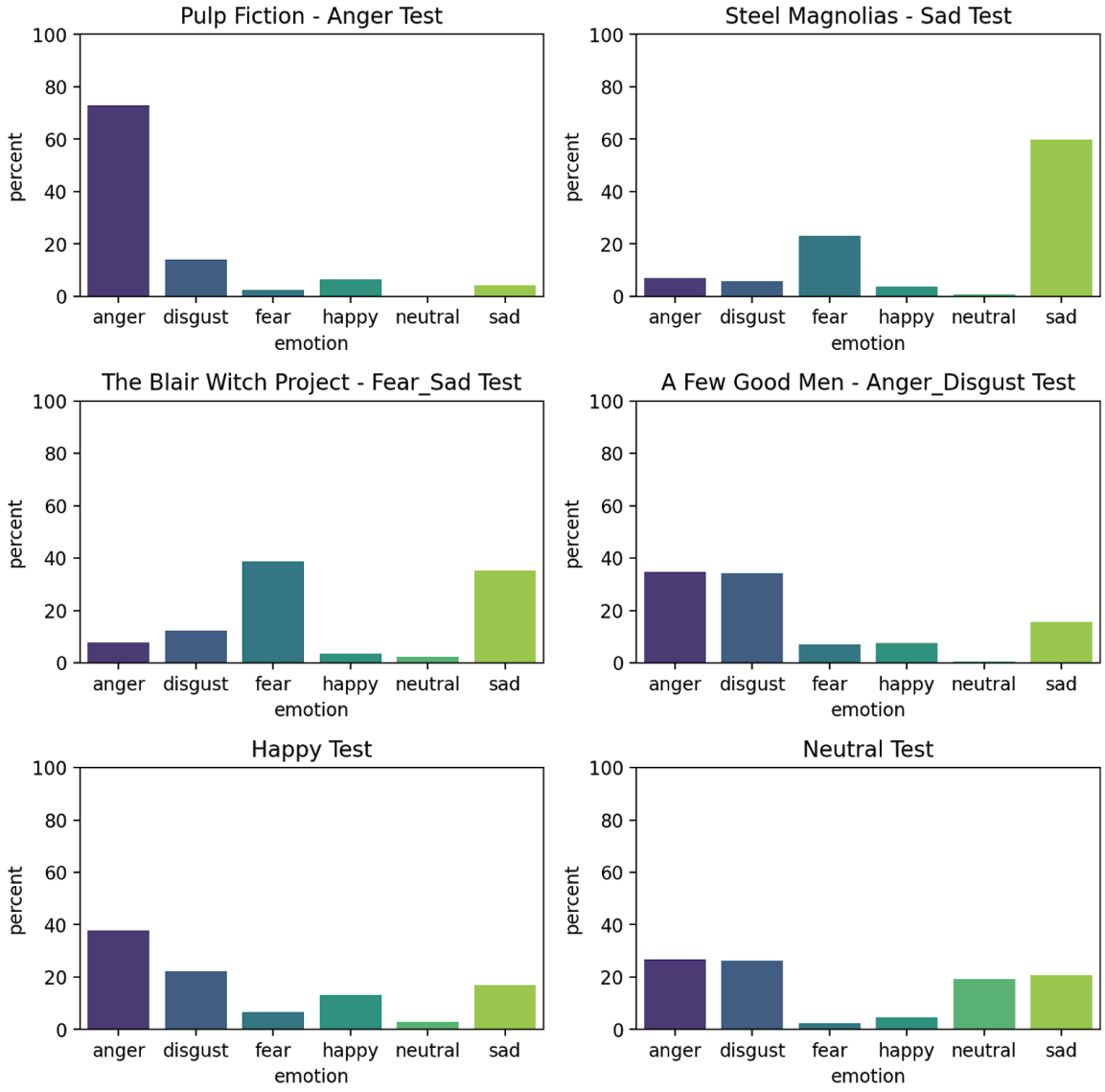
Model	Anger	Fear	Disgust	Sad	Happy	Neutral
AVG_EM	0.69	0.49	0.53	0.53	0.55	0.66
MAX_EM	0.68	0.43	0.54	0.51	0.54	0.64
AVG_SG	0.68	0.50	0.57	0.54	0.54	0.64
MAX_SG	0.69	0.48	0.47	0.53	0.49	0.62

9. Final Model

The two average pooling models were trained again using the whole dataset and their results were combined by averaging the emotion classification probability of both models. This joint model had a 100% training accuracy and was further tested using emotional movie scenes and my own recordings, six examples of which can be seen in Figure 10.

The movie scenes were selected under the requirement that they all had minimal background music and noise, and included at least ten seconds of dialogue by a single character, expressing the intended target emotion. I struggled to find movie scenes that fit this criteria for the labels **Happy** and **Neutral,** so instead I recorded myself reading Robert Frost's "Stopping by Woods on a Snowy Evening".

Figure 10: Bar graph showing the sum of all emotion probabilities.



Although this is very much a limited pool of test audio, the results do indicate the models overall performance. It appears that while the model was able to successfully determine **Anger, Sad, Fear,** and **Disgust,** it failed to consistently recognize **Happy** and **Neutral.**

It is also worth noting that in many cases, **Anger** and **Disgust** were often paired together, as were **Sad** and **Fear.** It was quite hard to find many movie clips that didn't have an overlap between these sets of emotions but whether that is more telling of model accuracy or the overlapping qualities of these emotions is unknown.

10. Conclusion

While the final model does not perform perfectly, its ability to recognize certain emotions from voices alone is still impressive. The final results show that the model is particularly keen on picking up negative emotions (**Anger, Fear, Sad, Disgust**), but struggles to identify the more positive ones (**Happy,Neutral**). This might be remedied by shifting the model to a binary classification format, but the current dataset would be unbalanced in favor of the negative emotions.

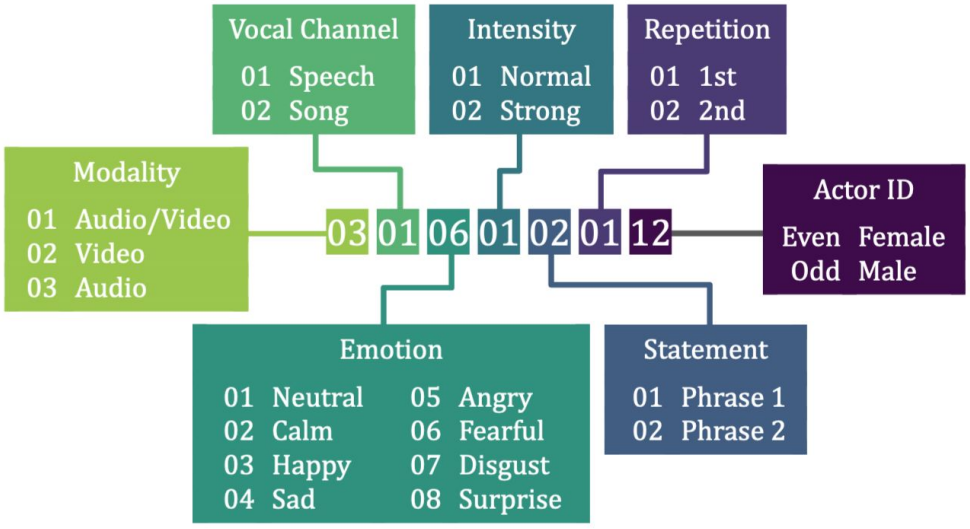
With more time and resources I would love to experiment with how adding sentiment analysis would affect these outcomes. Being able to add an understanding of the context and polarity of the words being said could greatly help the model pick out emotions, but may also confuse it further in particular cases, like sarcasm. It would also be interesting to test how this model actually compares to human detection of emotions, as even I struggled to determine the correct emotion labels for some of the training audio clips. It could be that a person would struggle just as much as the model in certain cases, and it would be interesting to compare the different results.

In the end, emotions are an incredibly complex and biological behavior that even we as people struggle to understand and correctly identify, especially without the guidance of facial features and body posture. While the model may fall short in certain aspects, it's quite fascinating that such a simple architecture and limited dataset managed to provide the accuracy it does.

Appendix

Figure I: Examples of filenames from both datasets.

a) *RAVDESS: a set of seven integer codes separated by dashes (i.e. 03-01-06-01-02-01-12.wav)*



b) *CREMA-D: a set of four codes separated by dashes (i.e. 1063_IEO_DIS_HI.wav)*

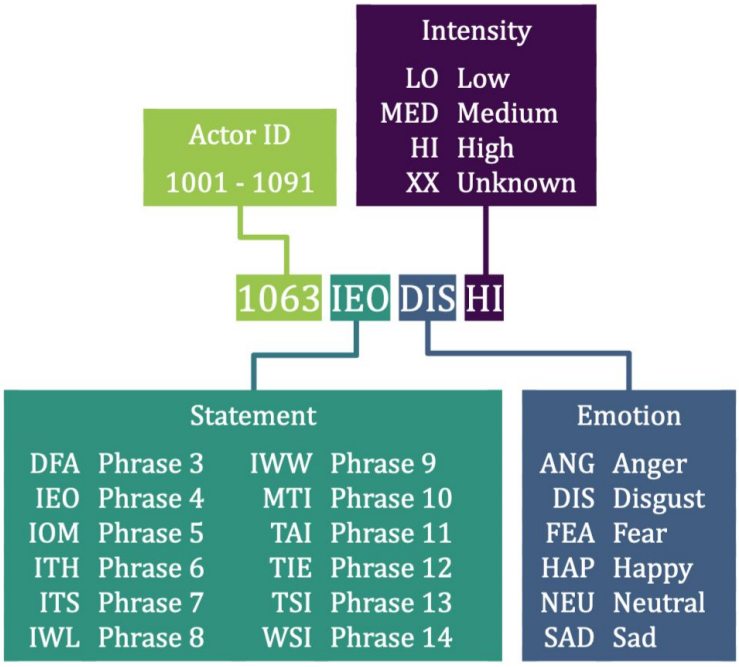


Figure II: Image of the mel-frequency spectrogram.

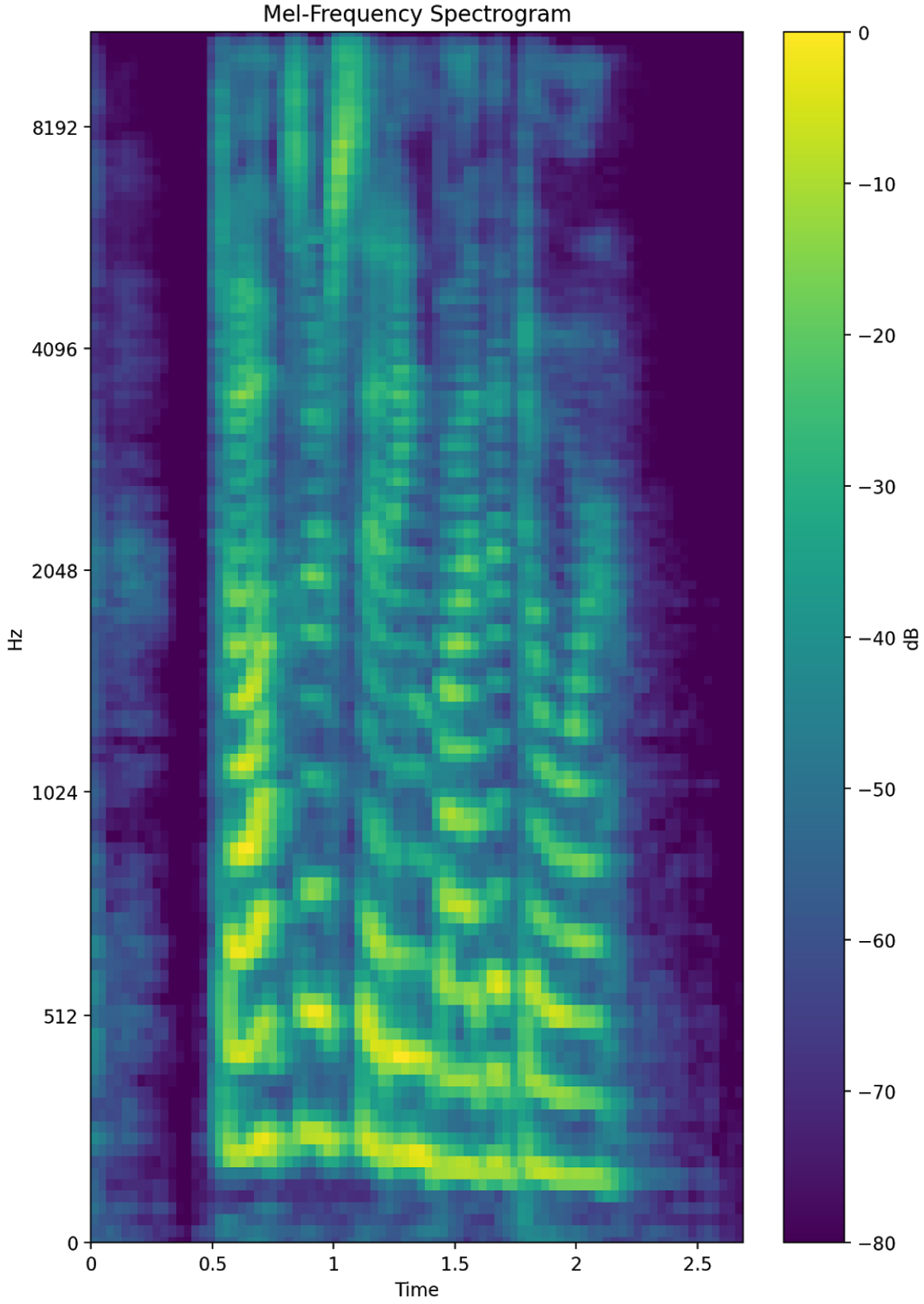


Figure III: Example of the final modeling image

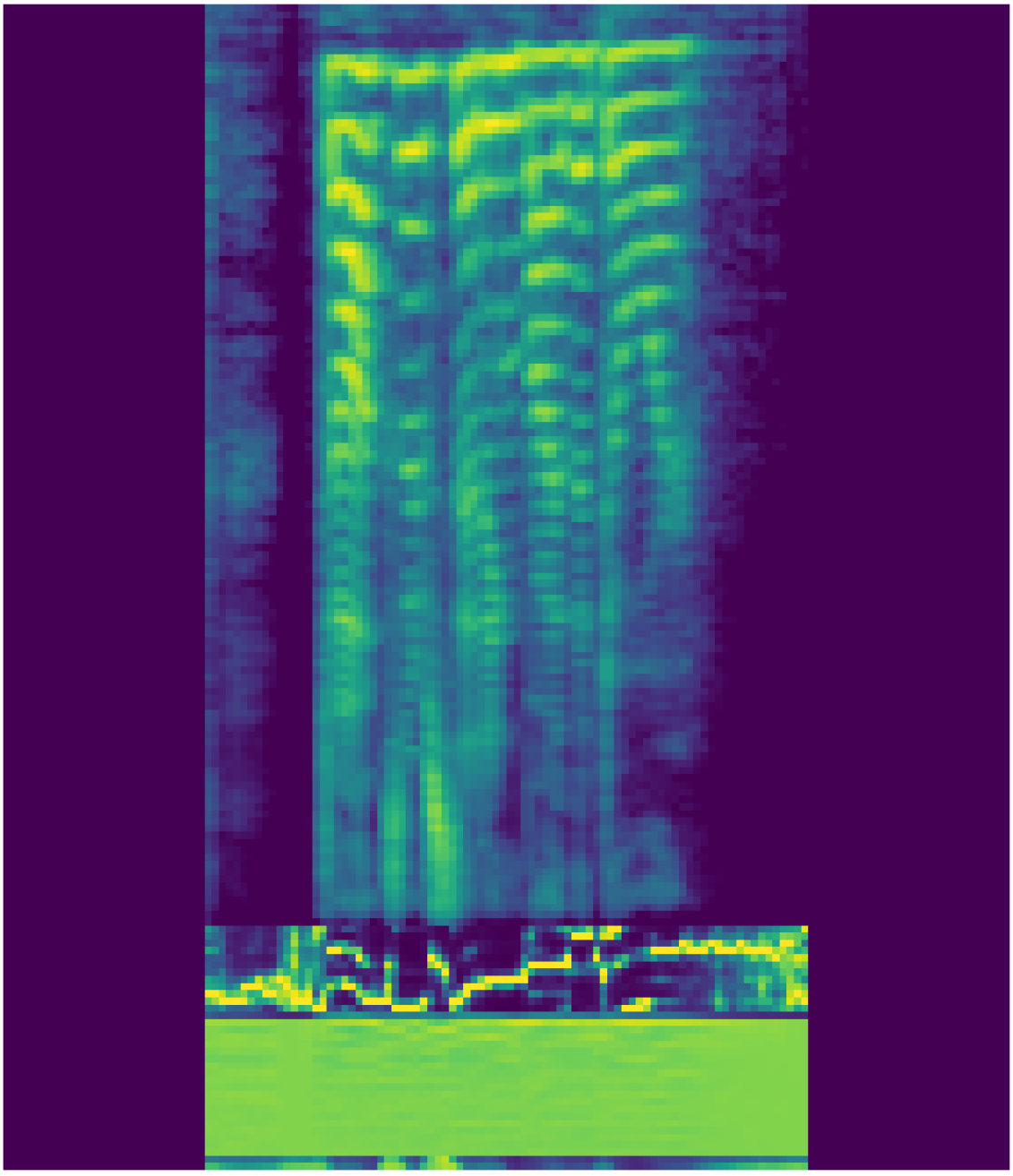


Figure IV: Full results for the models trained on both gender and emotion.