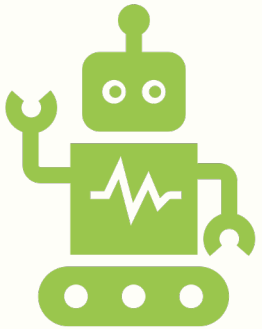
Voice Emotion Analysis

Overarching Question

Can a machine recognize the emotions in a person's voice?



Can a machine recognize the emotions in a person's voice?

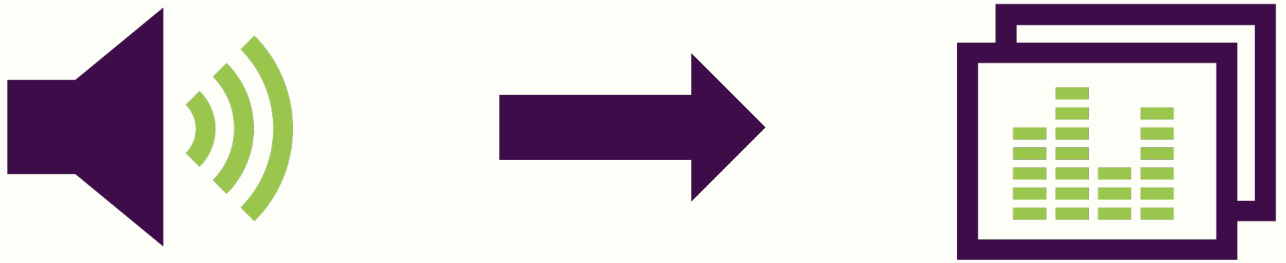
Why Does It Matter?

- Voice recognition technology is used daily all over the world.
  - Digital assistants, automated phone systems, hands free phone pairing
- Adding the emotional context of voices could greatly increase a computers understanding of what's being said.

Can a machine recognize the emotions in a person's voice?

How Can It Be Done?

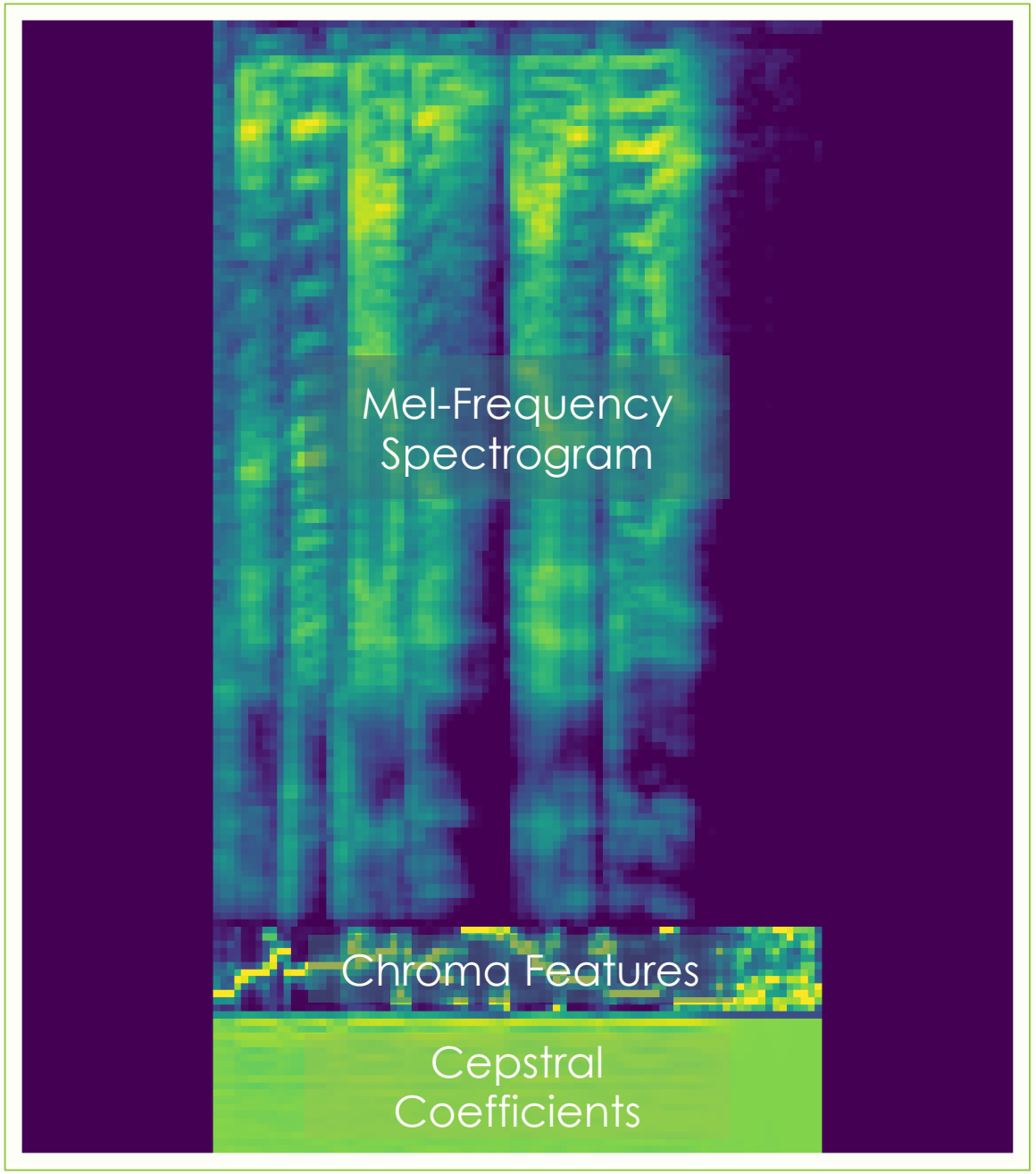
- Turn one-dimensional sound into a two-dimensional picture.
  - They say a pictures worth a thousand words! Extracting audio features as images can provide a much greater source of information.
- Use these images to train a convolutional neutral network (CNN).
  - CNN Models are the go-to for analyzing images, best known for classifying handwriting and even cat pics.



Feature Selection

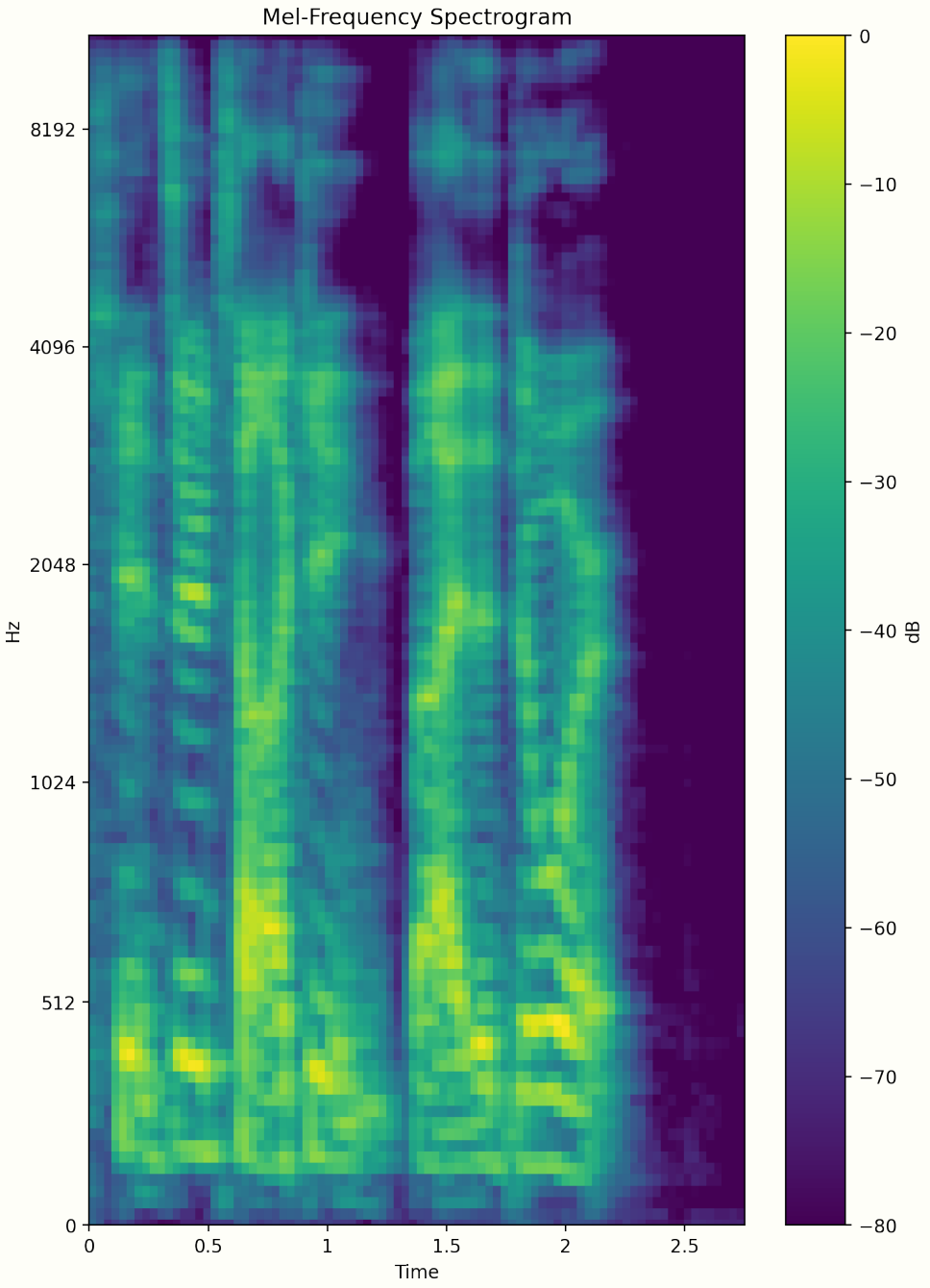
Feature Selection

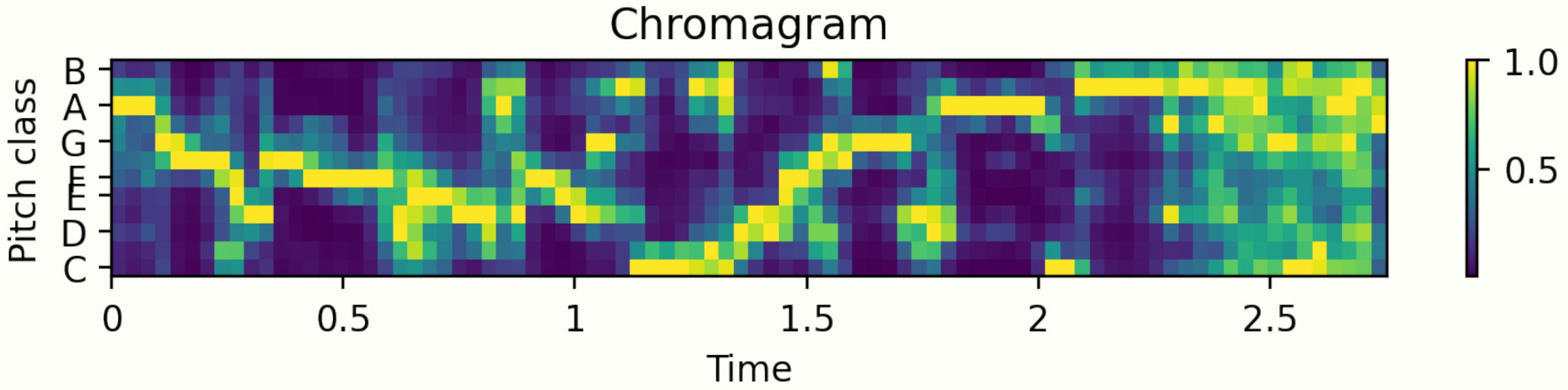
- Each audio clip was broken up into three features.
- These features were extracted as 2D arrays and stacked together to make a single image for each audio clip.



Mel-Frequency Spectrogram

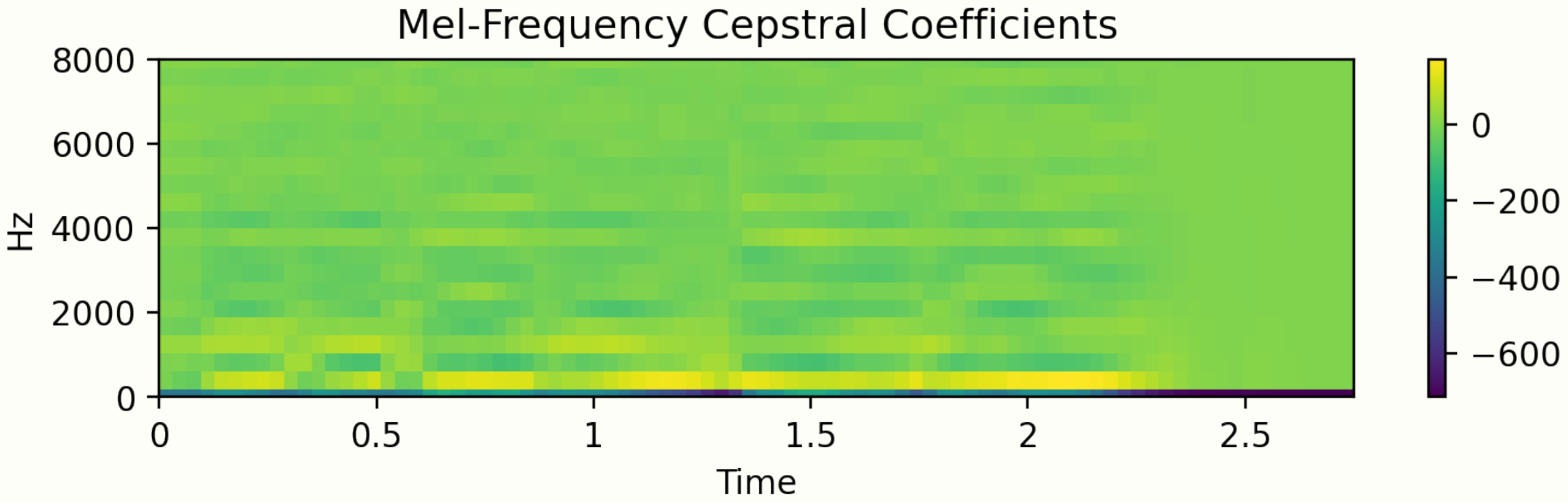
- A visualization of the frequency spectrum created using the mel scale.
  - Mel Scale: non-linear pitch scale based on human hearing
- Shows how the energy of a signal changes with time, focusing particularly on the human perception of these characteristics.





Chroma Features

- A break down the energy of an audio signal into one of the twelve pitch classes.
- Provides an overall profile of how the pitch of a sound shifts with time.



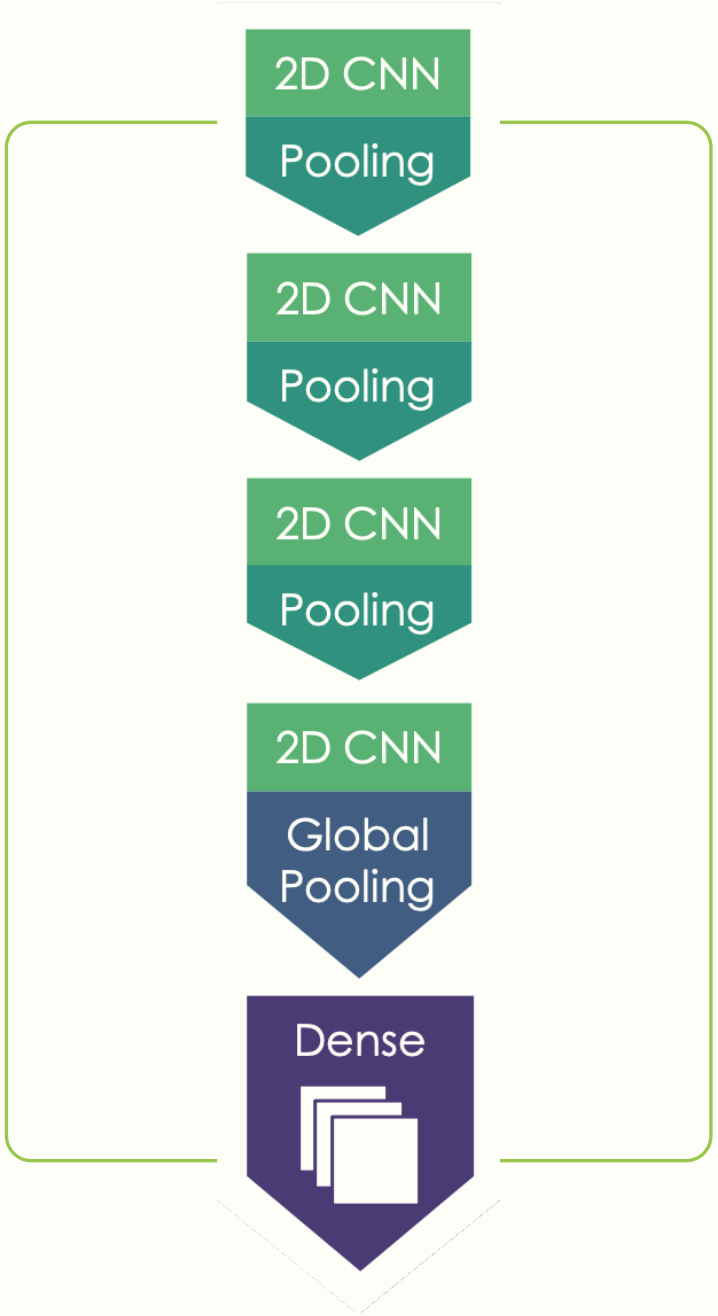
Cepstral Coefficients

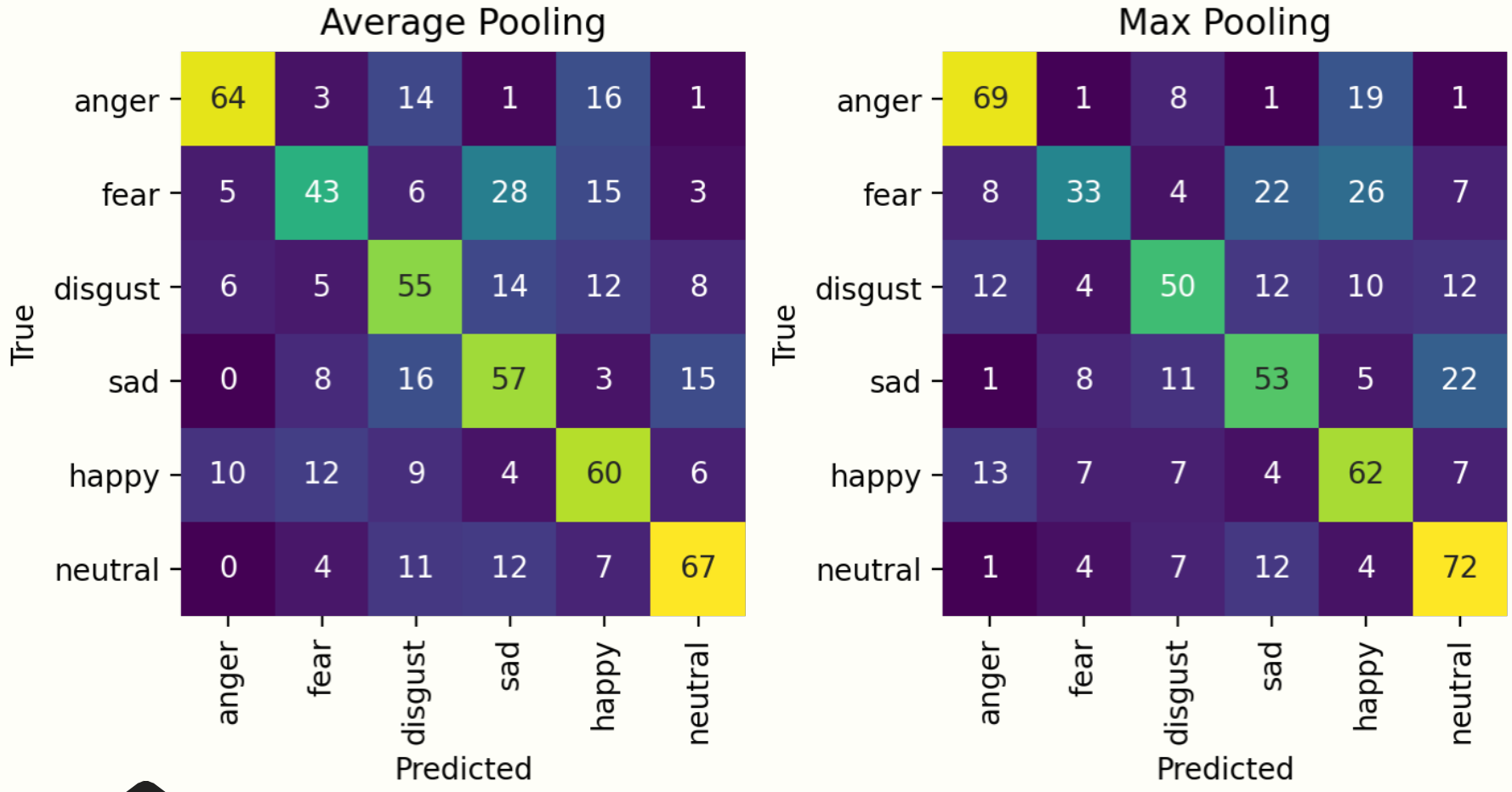
- Coefficients of the power spectrum of the audio.
- Typically used to model and understand the qualities of human voices.


Model

Four Models

- Two model architectures were tested, each using a different pooling layer between convolution layers.
  - **MAX**: max pooling layers
  - **AVG**: average pooling layers
- Both were trained twice with using two sets of labels
  - **EM**: 6 emotion labels
  - **SG**: 12 emotion/gender labels



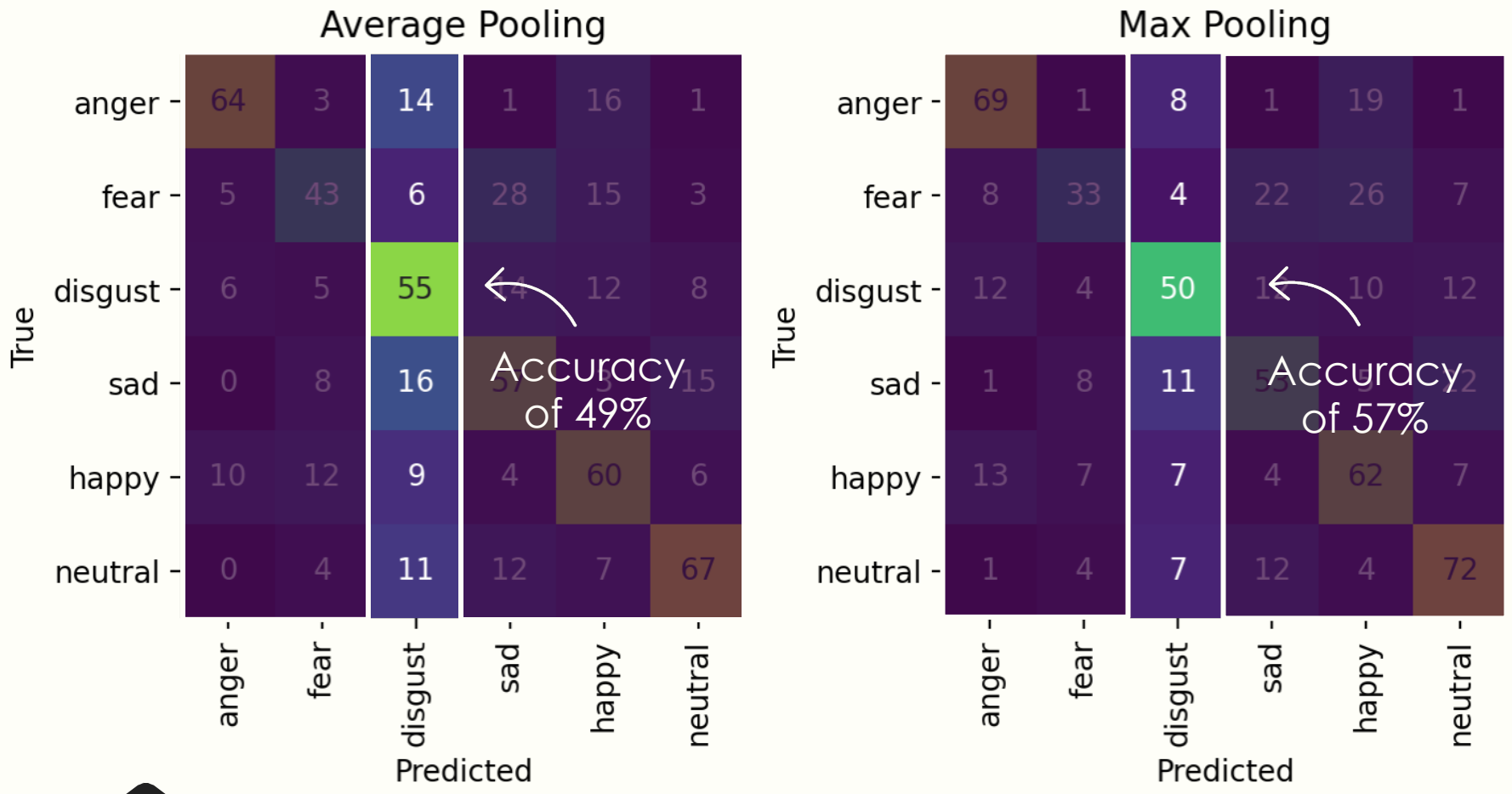


Model Results

Emotion

- Accuracy of models trained on emotion labels only:
  - AVG_EM: **57%**
  - MAX_EM: **56%**

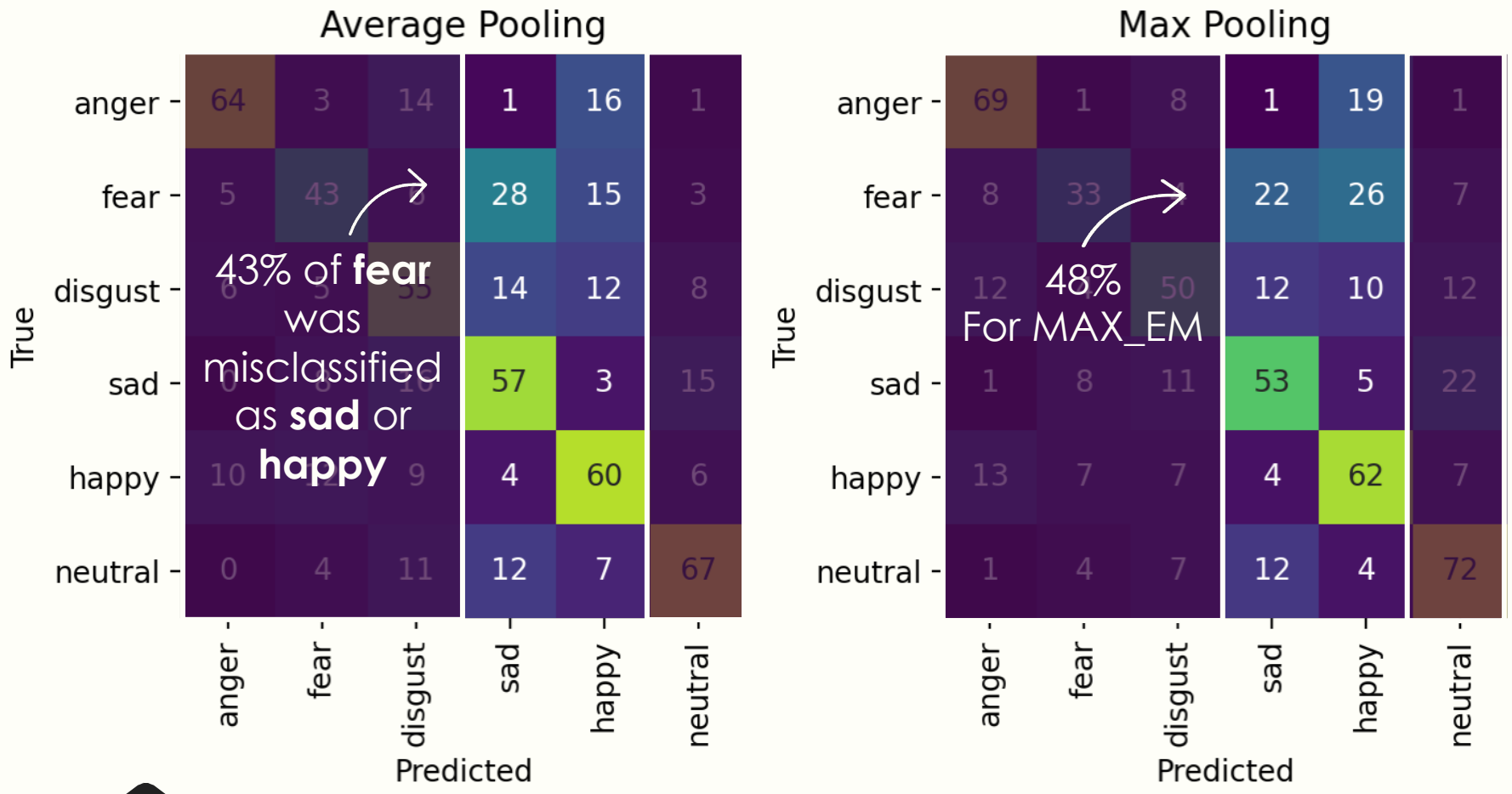
*** random chance: 17%**



Model Results

Emotion

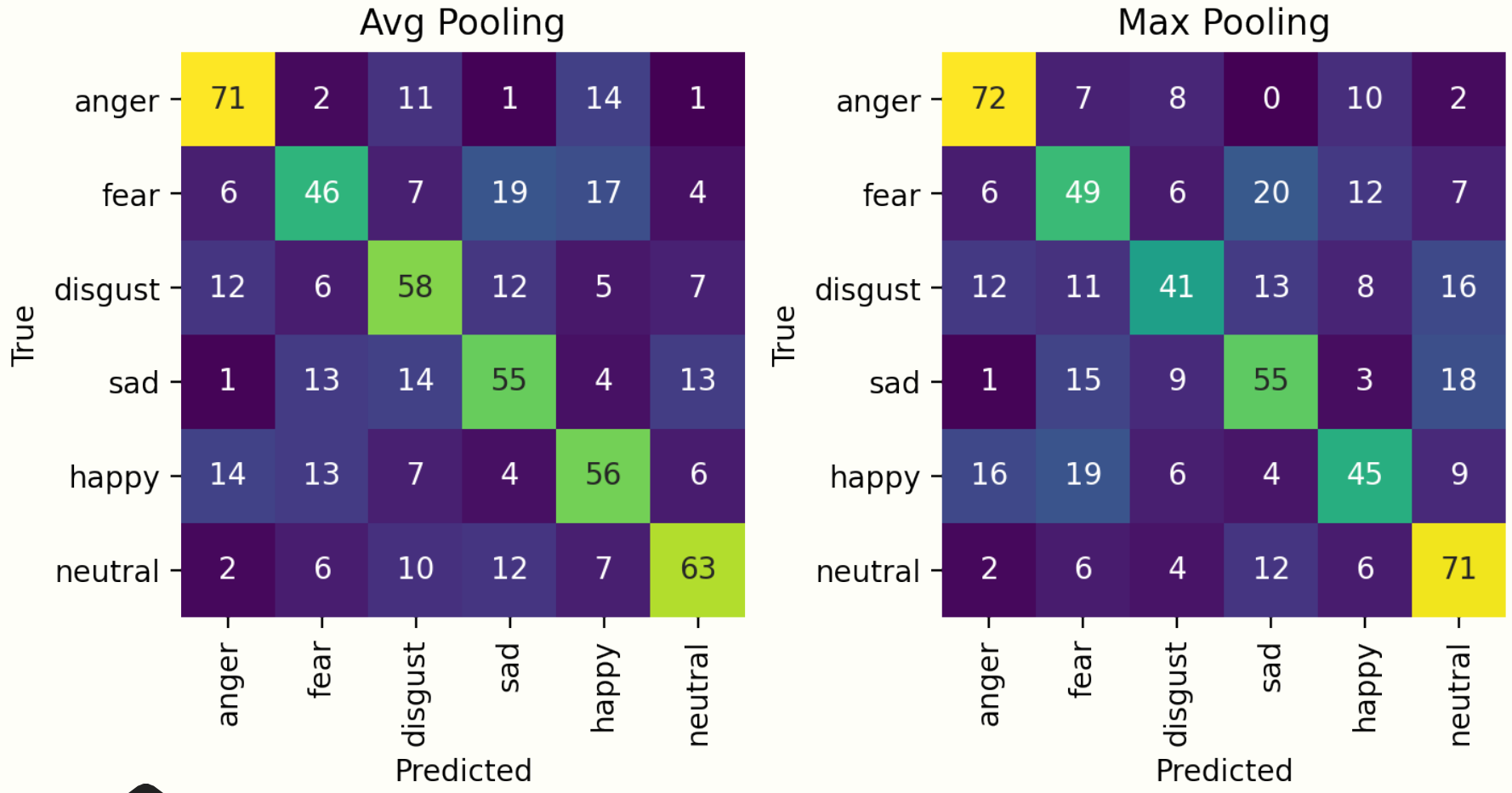
- AVG_EM had slightly better scores overall…
  - But MAX_EM was better at classifying **disgust.**



Model Results

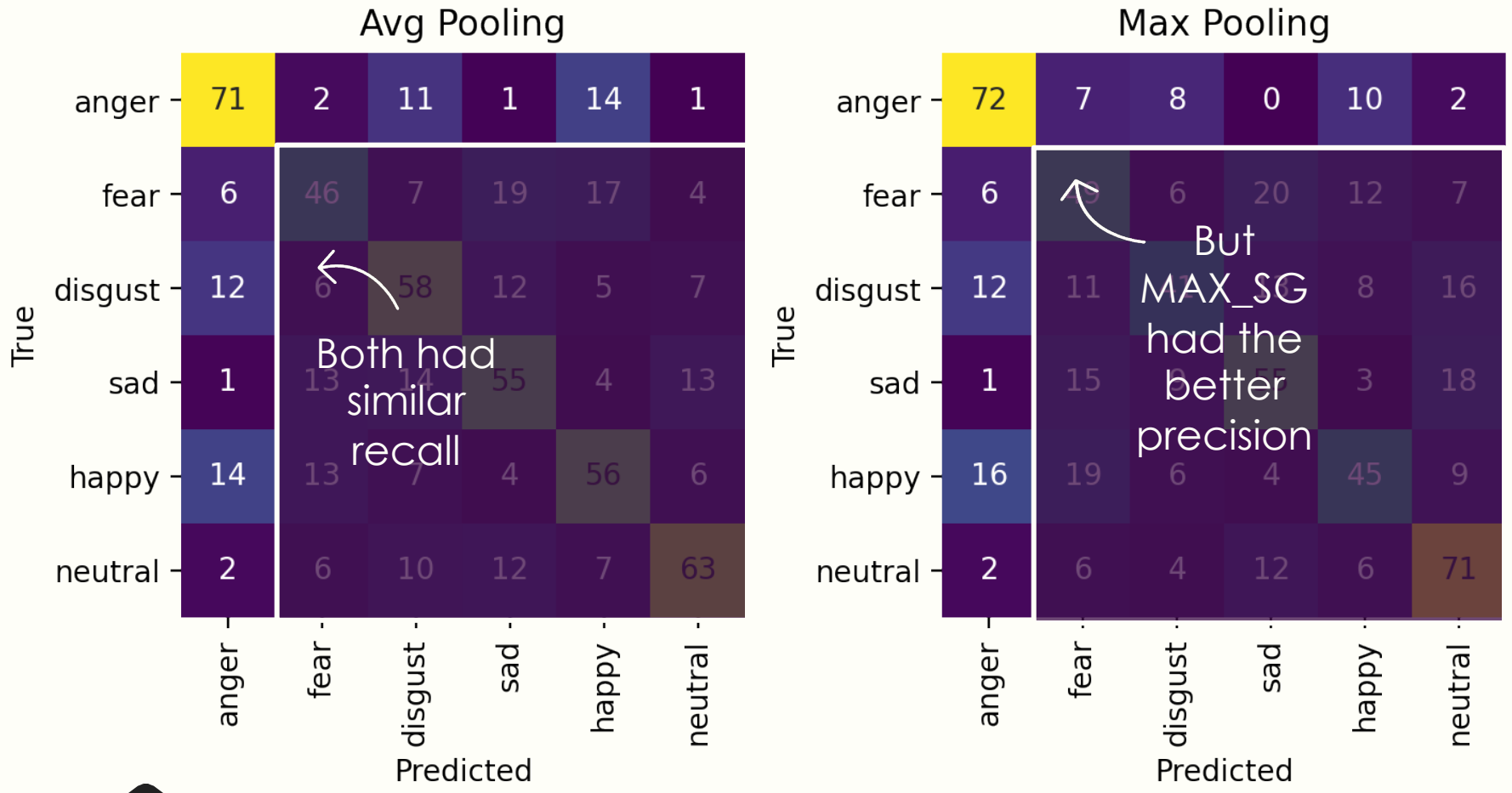
Emotion

- Both had a tendency to over-label emotions as either **sad** or **happy**…
  - Particularly **fear**



Model Results
Emotion & Gender

- Accuracy of models trained on emotion and gender:
  - AVG_SG: **58%**
  - MAX_SG: **55%**



Model Results
Emotion & Gender

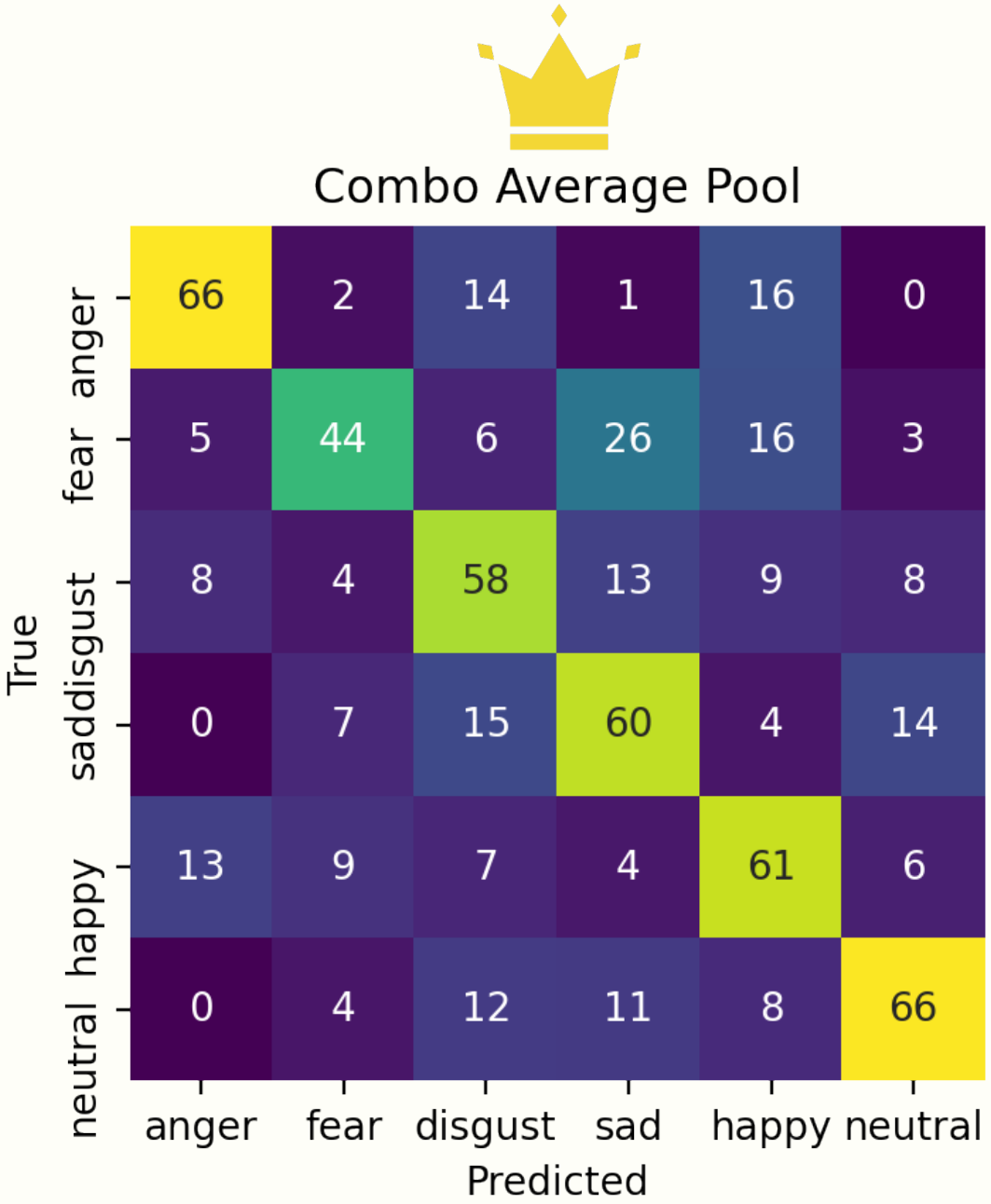
- AVG_SG was overall a much better model than MAX_SG…
  - Although MAX_SG was slightly better at determining **anger.**

Model Results

- AVG_SG was the best model overall.
- In terms of the independent emotions though, the best scores were split between AVG_EM and AVG_SG.

So what if the two best models were combined?

Model	ANG	FEA	DIS	SAD	HAP	NEU
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

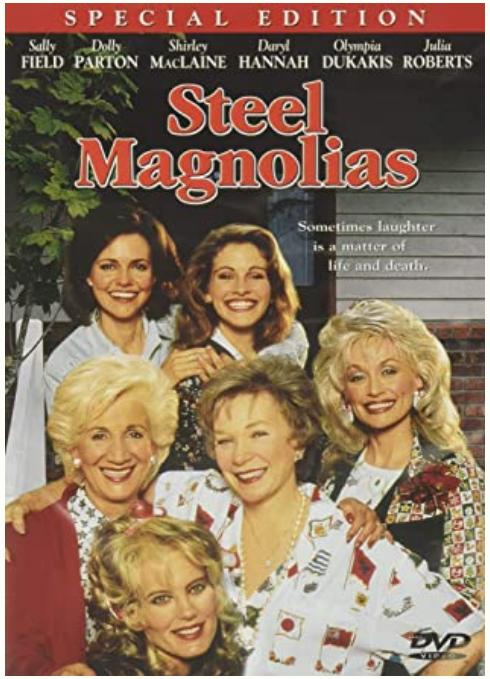


Combination Model

- Accuracy of **59%**
- Scores highest of all the models for all emotions except for **disgust**.
  - F1 score beat by AVG_SG by only **0.01**

Overall Best Model!

Movie Scenes!



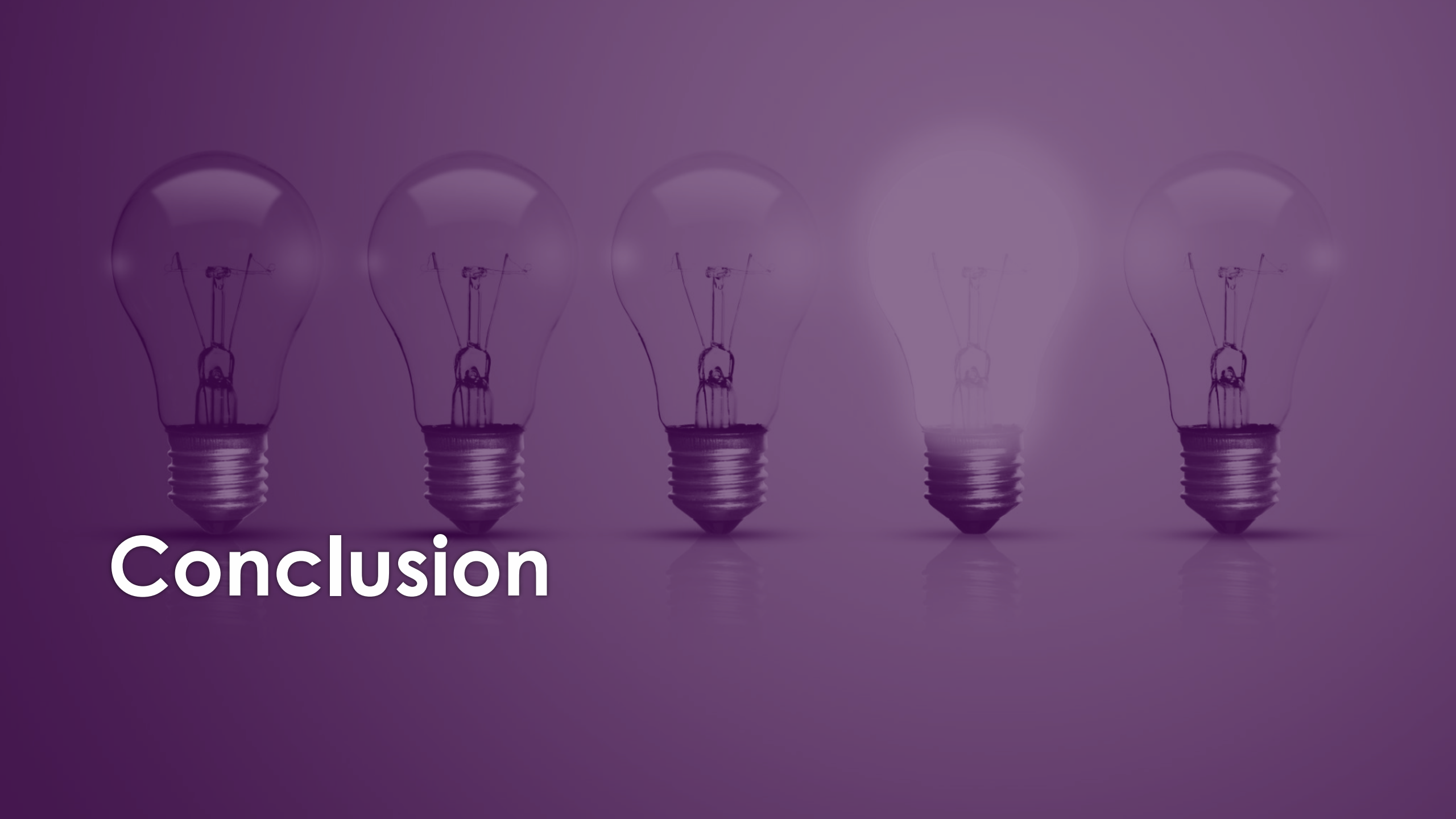
*** And a couple of my own recordings**

Final Model Testing

- The final combination model was trained on all the audio clips in the dataset.
- With no more data to test on, how can we find out how well it preforms?

Final Model Testing

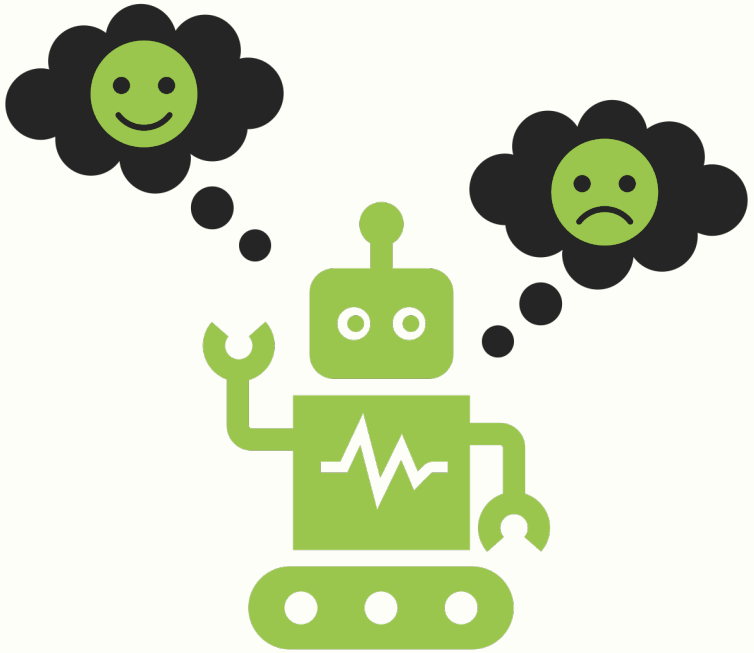
- Overall results:
  - The final model was good at classifying negative emotions
  - But struggled to classify the more positive emotions.
    - Mislabeling them as **anger** and **disgust**



Conclusion

Summary

- A combination of the average pooling models gave the best test results (59%)
- The final model could identify the negative emotions **anger**, **sad**, **disgust**, and **fear,** but struggled with the more positive emotions **happy** and **neutral.**
  - This may be because these emotions tend to be less intensive
  - Neutral can be seen as an ‘absence’ of emotion rather than an emotion itself.



Moving Forwards

There's always room for improvement! What else can be done?

- **More Data**
  - A greater pool of audio clips to train on can help the model learn.
- **Higher Complexity**
  - This model was made simple to save resources, a more complex model may provide higher accuracy.
- **Binary Classification**
  - The model was good at identifying negative emotions, perhaps it would be easier if the model was trained to classify emotions simply as either positive or negative.

Can a machine recognize the emotions in a person's voice?

Yes!

- Even with such limited data and resources, this simple model was able to identify emotions quite well.
- While currently it has a limited accuracy, there are plenty of directions to pursue continued growth and improvement!