Sound Classification

Cough, Laugh, Sigh, Sneeze, Sniff, Throat Clear

John Guo

Motivation

Speech recognition is all around us:

- Siri, Alexa, Google Home
- Youtube or Zoom captioning
- Customer service...
- And many more (wiki/Speech_recognition)

Motivation

Speech recognition is all around us:

- Siri, Alexa, Google Home
- Youtube or Zoom captioning
- **Customer service...**
- And many more (wiki/Speech_recognition)

Problem Statement

- Automatic speech recognition (ASR) is hard
  - people make tons of noises unrelated to speech
  - cough, laugh, sigh, sneeze, sniff, throat clear

Problem Statement

- Automatic speech recognition (ASR) is hard
  - people make tons of noises unrelated to speech
  - cough, laugh, sigh, sneeze, sniff, throat clear

- Goal is to train a neural network that takes in short audio files and classifies it as one of the noises listed above (success: accuracy $> 90\%$)

Problem Statement

- Automatic speech recognition (ASR) is hard
  - people make tons of noises unrelated to speech
  - **cough, laugh, sigh, sneeze, sniff, throat clear**

- Goal is to train a neural network that takes in short audio files and classifies it as one of the noises listed above (success: accuracy $> 90\%$)

Problem Statement

- Automatic speech recognition (ASR) is hard
  - people make tons of noises unrelated to speech
  - **cough, laugh, sigh, sneeze, sniff, throat clear**

- Goal is to train a neural network that takes in short audio files and classifies it as one of the noises listed above (success: accuracy $> 90\%$)

- Hope is that this could eventually be paired with ASR

About our data

- ~21,000 audio files
- ~ 3,000 contributors
- Typically 2-7 seconds long
- Included some info about contributors:
  - Age
  - Gender
  - Country
  - Language

Cough

Laugh

Sigh

Sneeze

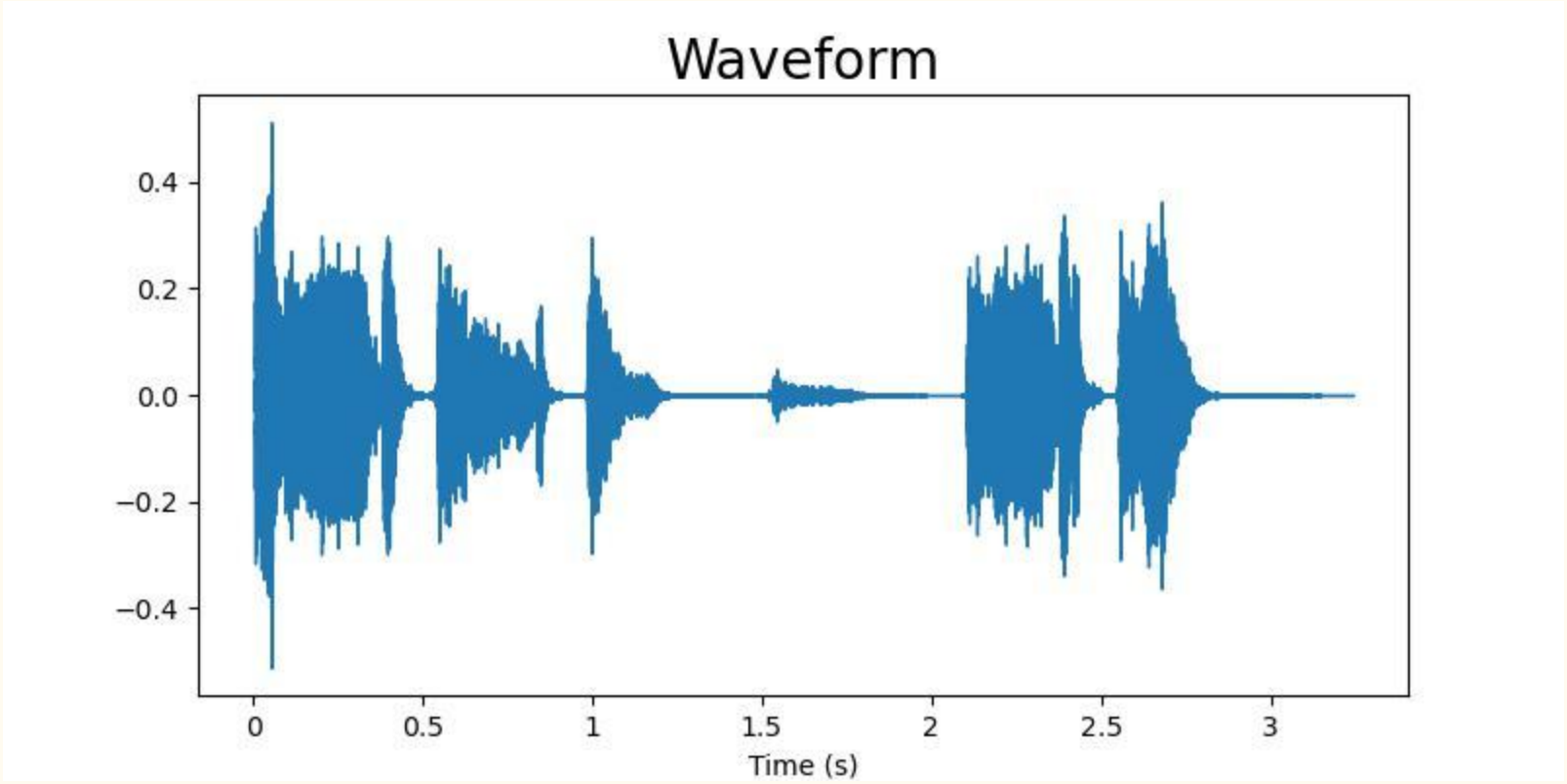
Sniff

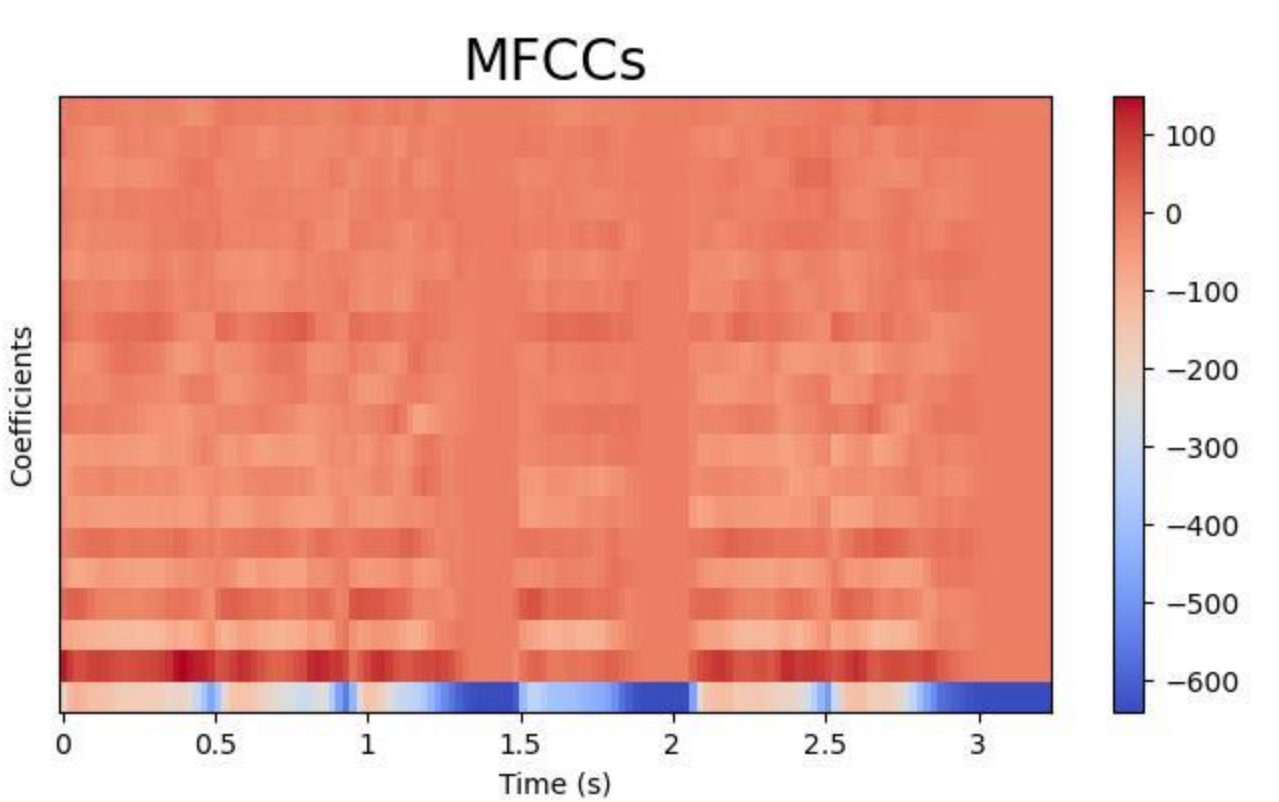
Throat Clear

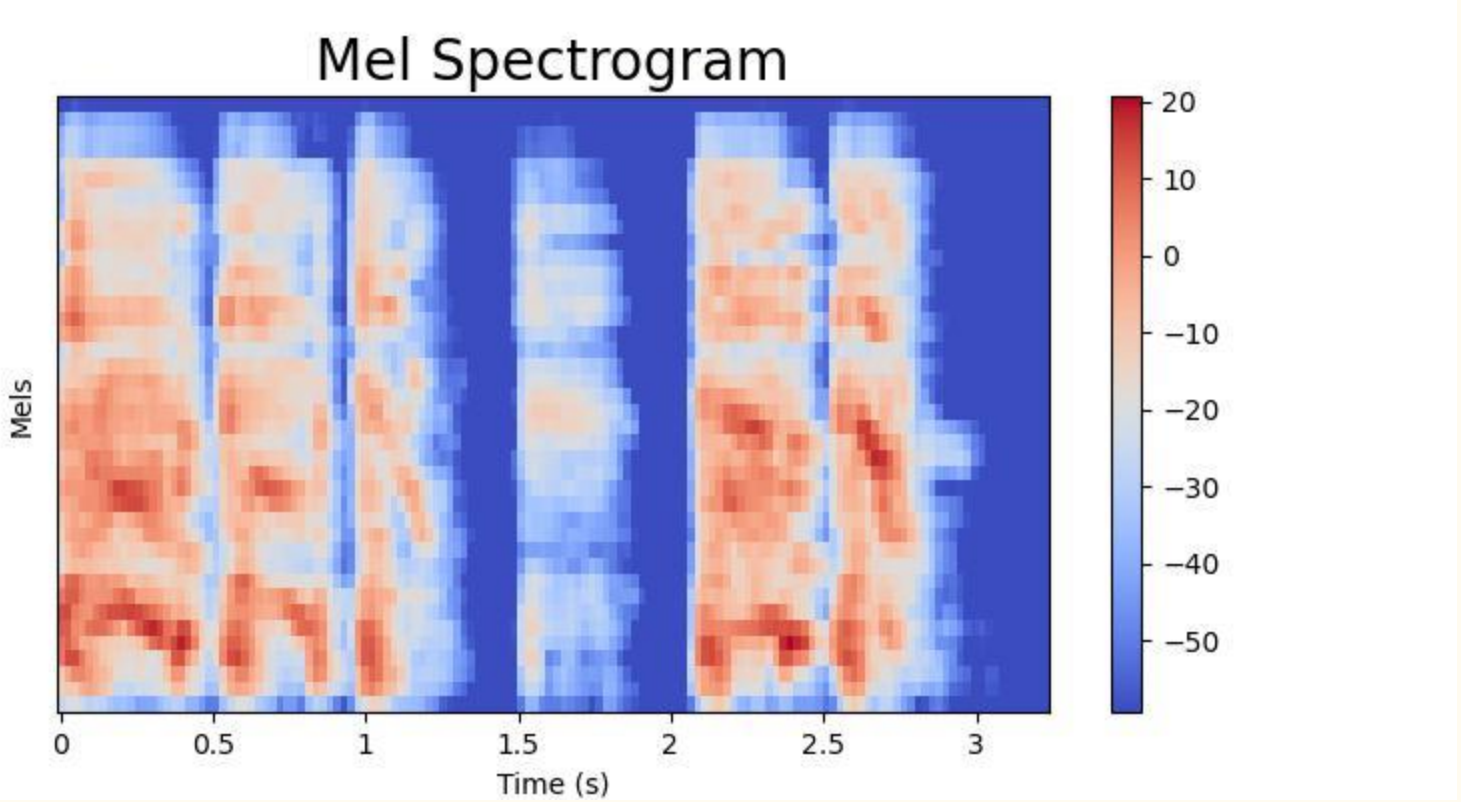
I don’t think that means what you think it means...

Labeled as a sigh:

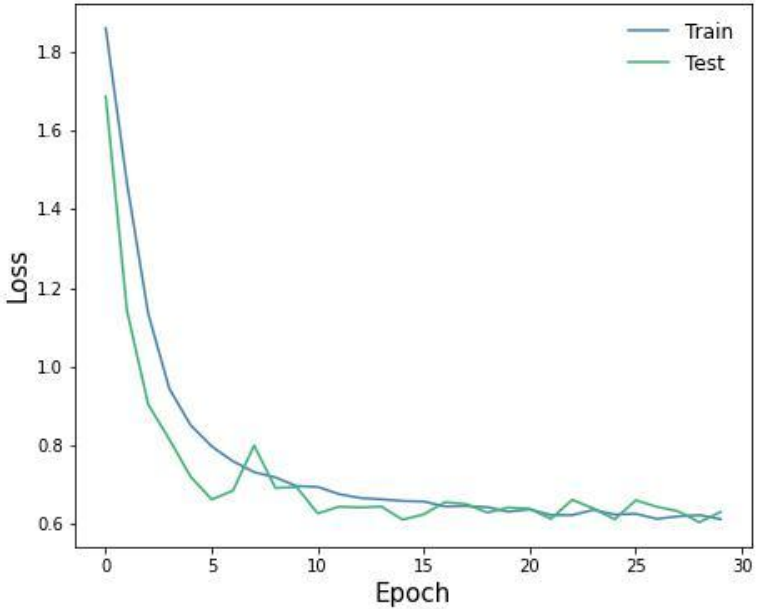
Image provided by Vecteezy

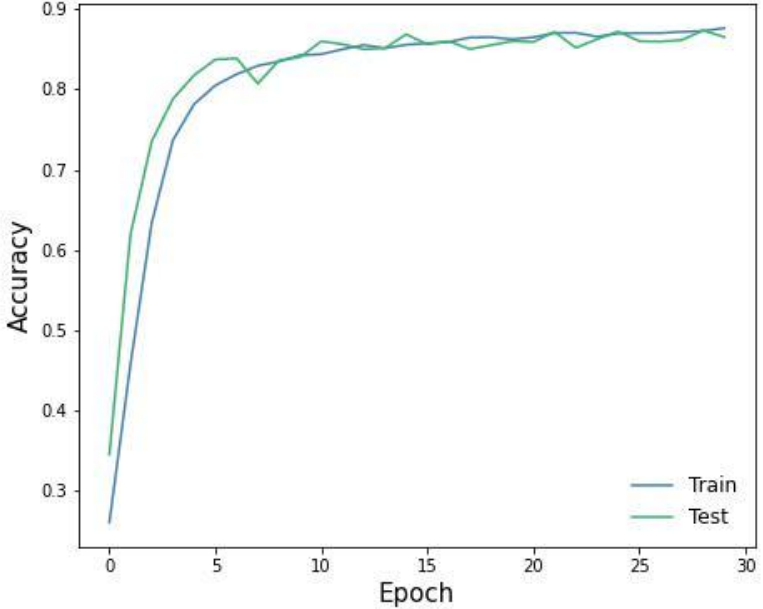

Waveform
0.4
0.2
0.0
−0.2
−0.4
0
0.5
1
1.5
2
2.5
3
Time (s)


MFCCs
Coefficients
Time (s)
0
0.5
1
1.5
2
2.5
3
100
0
−100
−200
−300
−400
−500
−600


Mel Spectrogram
Mels
0
0.5
1
1.5
2
2.5
3
Time (s)
20
10
0
−10
−20
−30
−40
−50

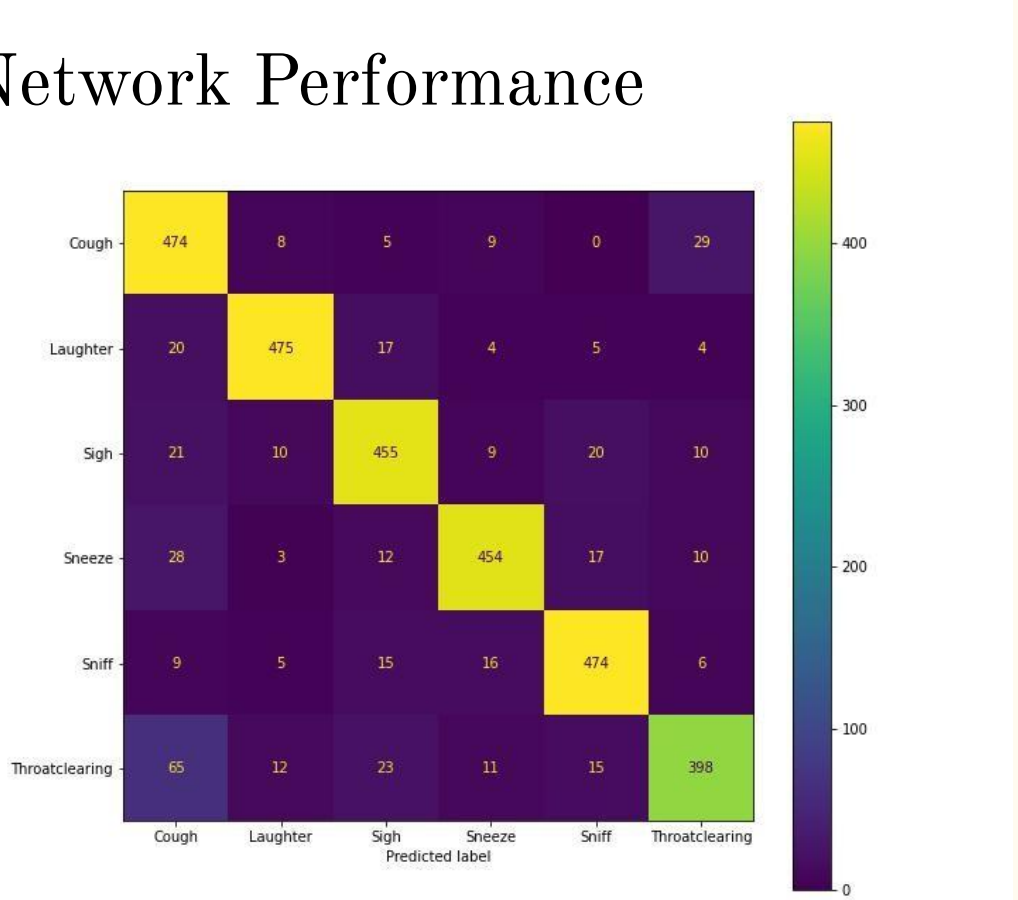
MFCC CNN Model Performance





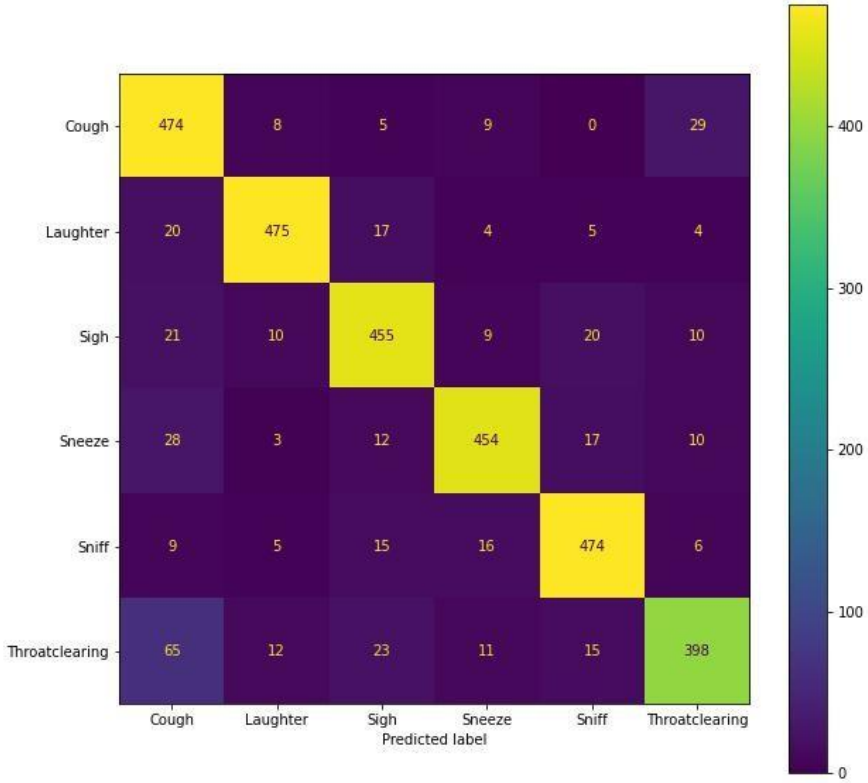
MFCC Conv. Neural Network Performance

- Big numbers on diagonal



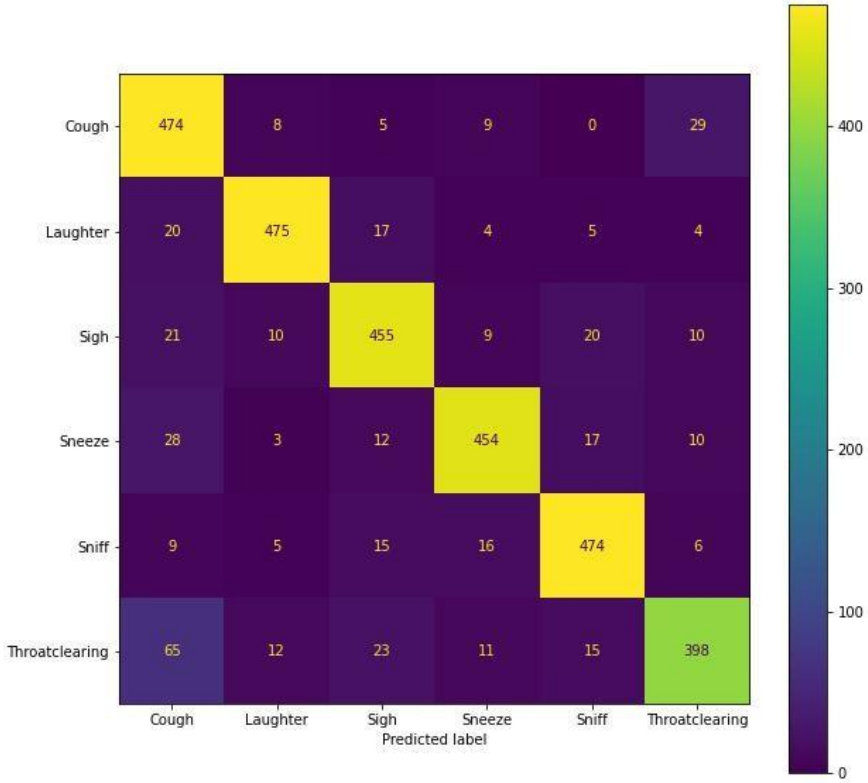
MFCC Conv. Neural Network Performance

- Big numbers on diagonal
  - **Good!**



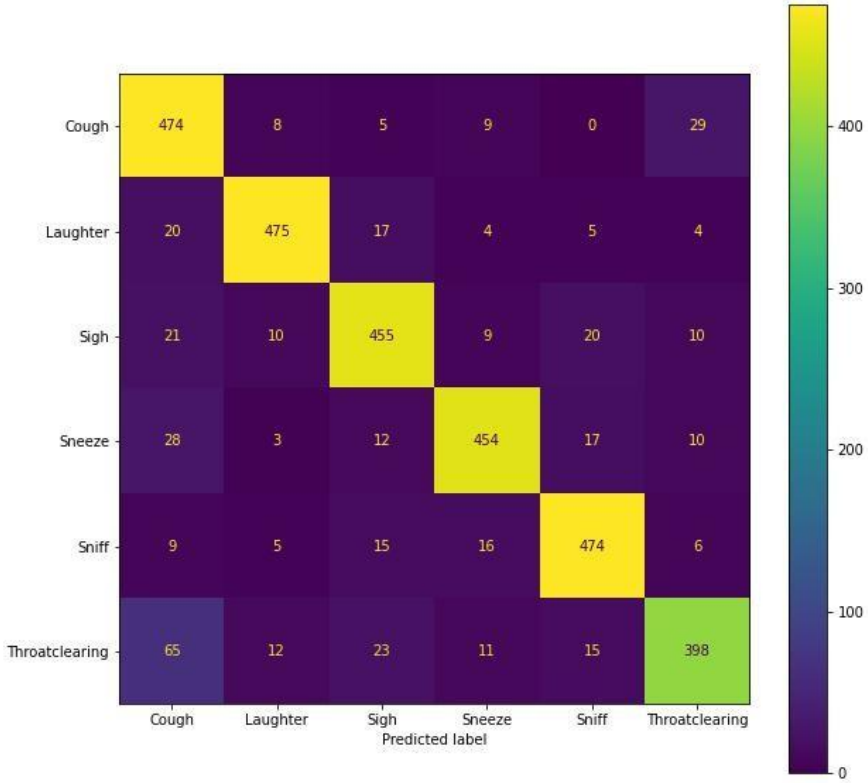
MFCC Conv. Neural Network Performance

- Big numbers on diagonal
  - **Good!**
- Small numbers off diagonal



MFCC Conv. Neural Network Performance

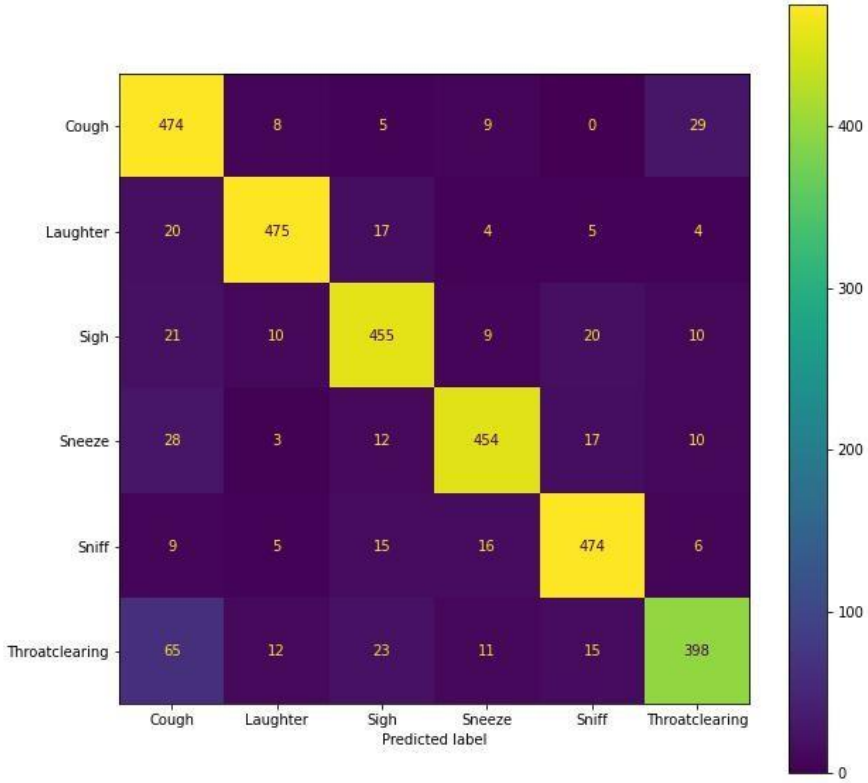
- Big numbers on diagonal
  - **Good!**
- Small numbers off diagonal
  - **Good!**



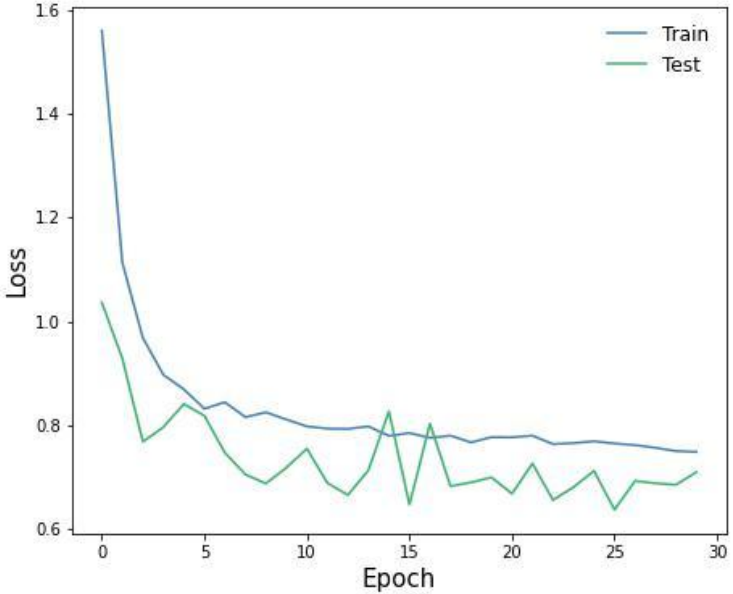
MFCC Conv. Neural Network Performance

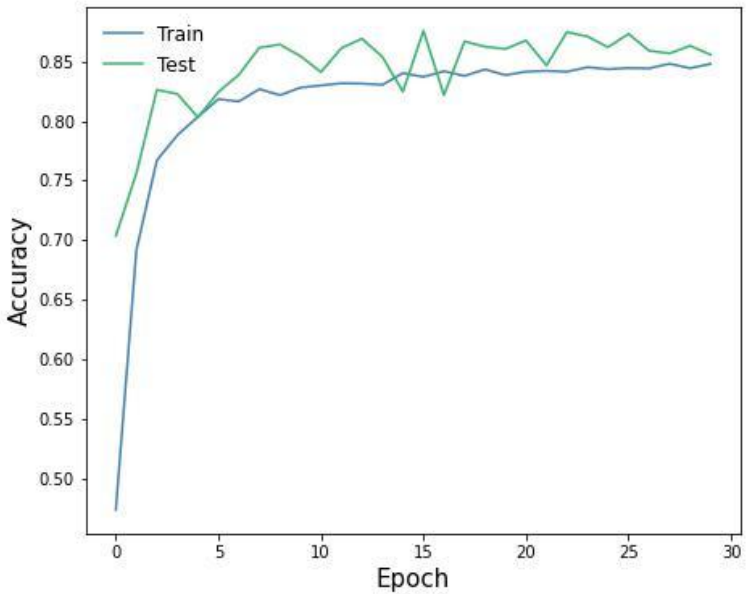
- Big numbers on diagonal
  - **Good!**
- Small numbers off diagonal
  - **Good!**

- Labels some throat clears as coughs



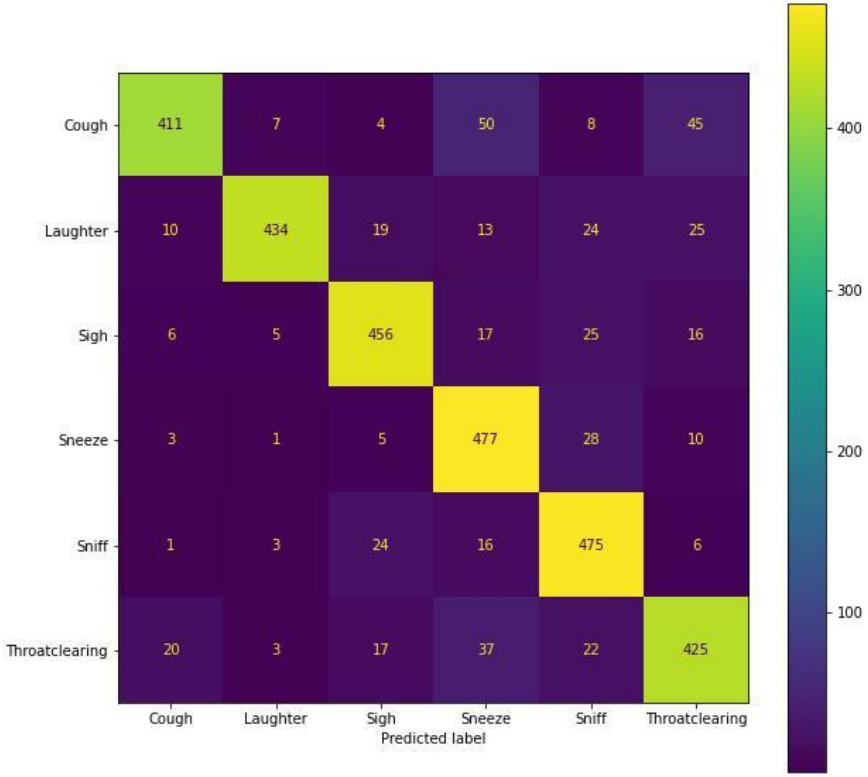
Mel Spectrogram CNN Model Performance





Mel Spectrogram Conv. Neural Network Performance

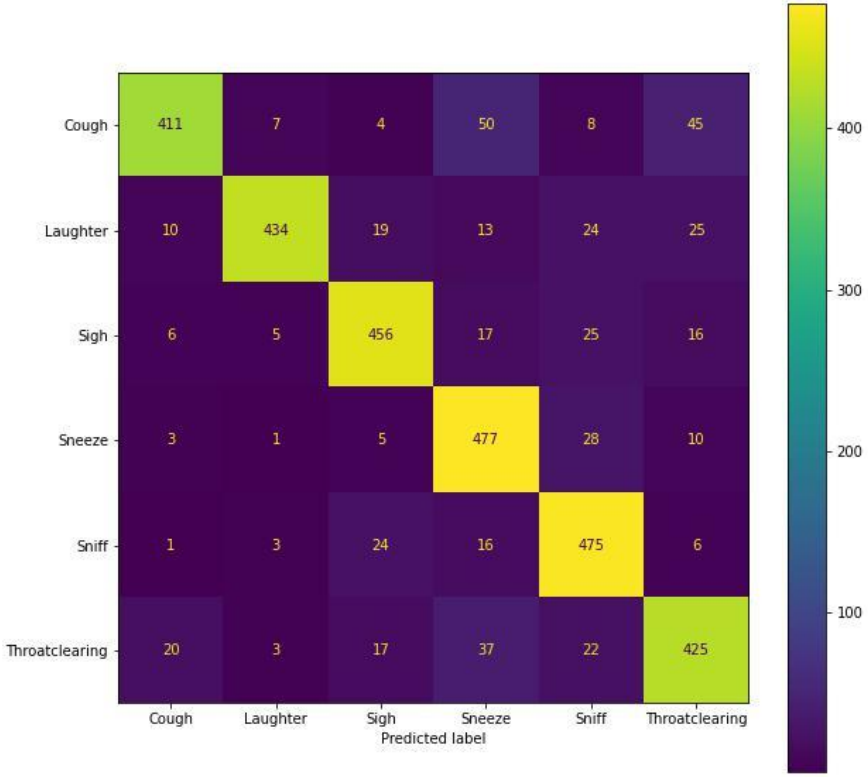
- Big numbers on diagonal
  - **Good!**
- Small numbers off diagonal
  - **Good!**



Mel Spectrogram Conv. Neural Network Performance

- Big numbers on diagonal
  - **Good!**
- Small numbers off diagonal
  - **Good!**

- Has more general trouble:
  - Sneeze
  - Sniff
  - Throat clearing



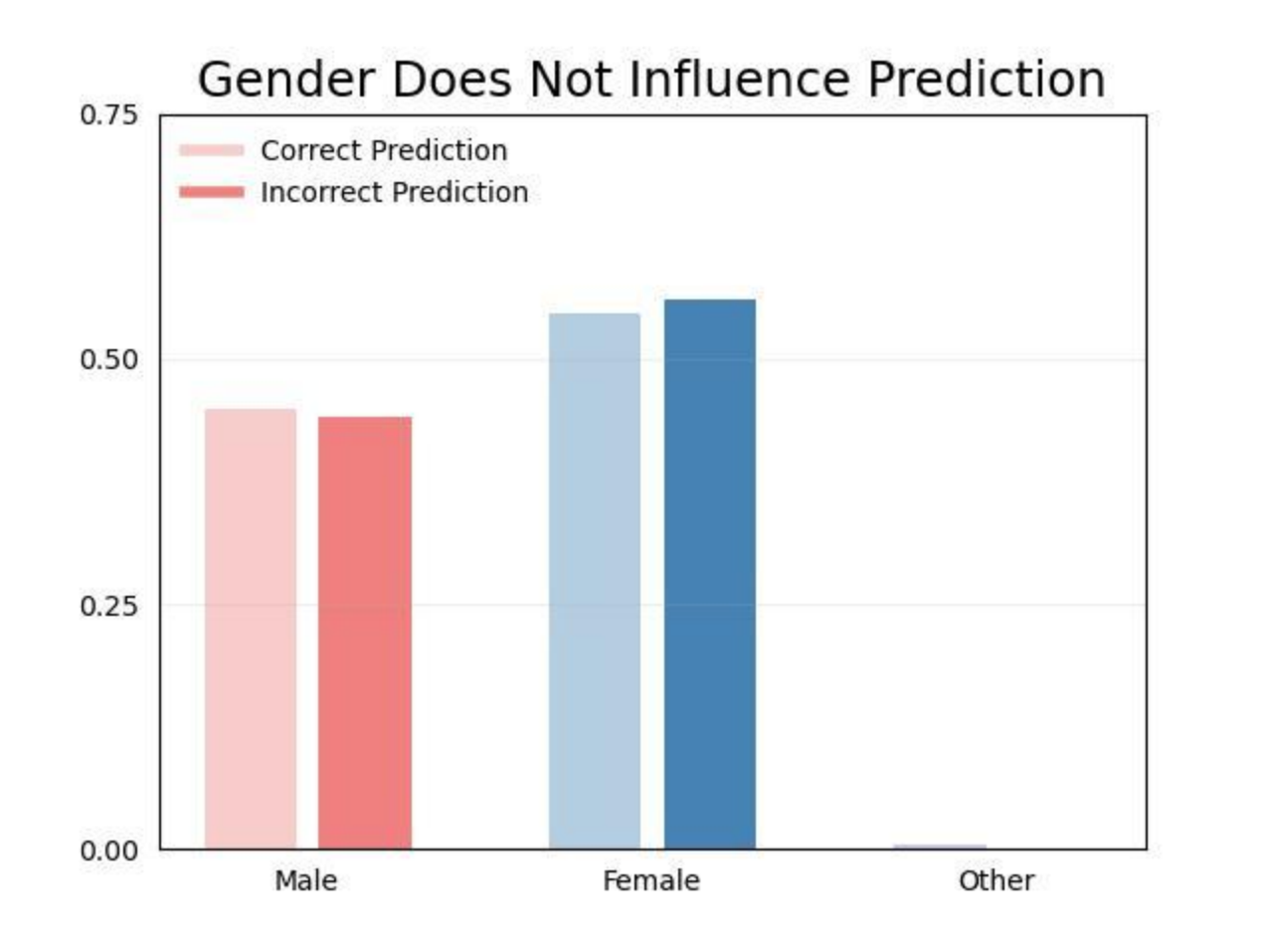
Our production model will be...

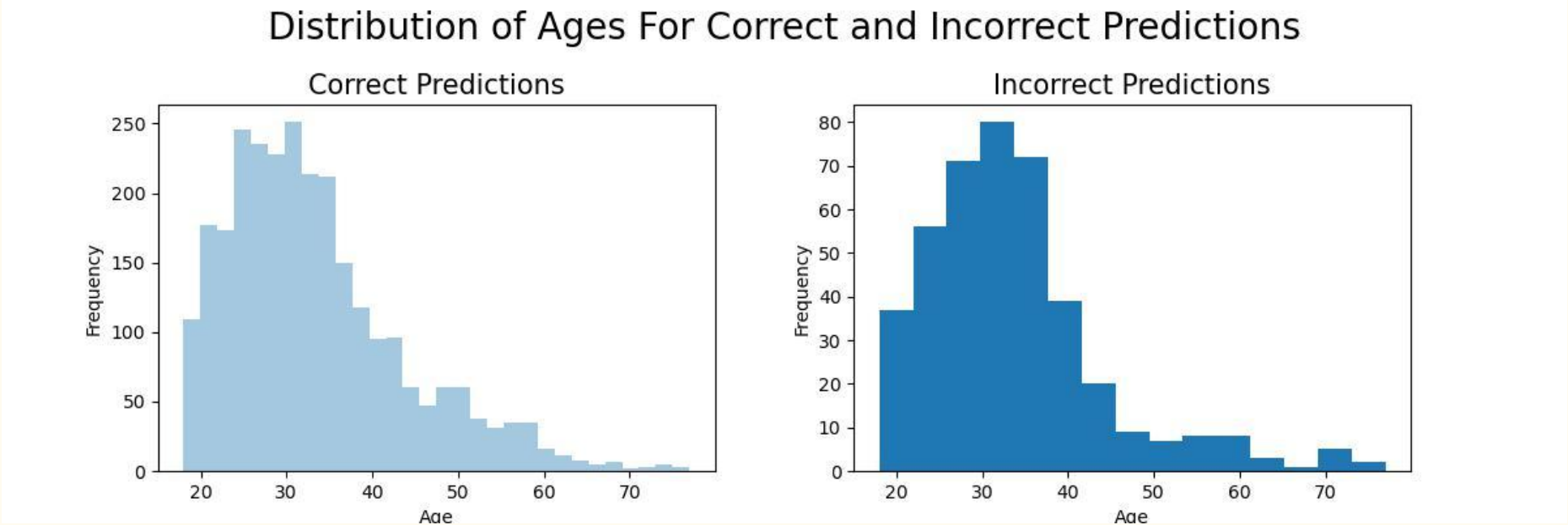
Our production model will be...

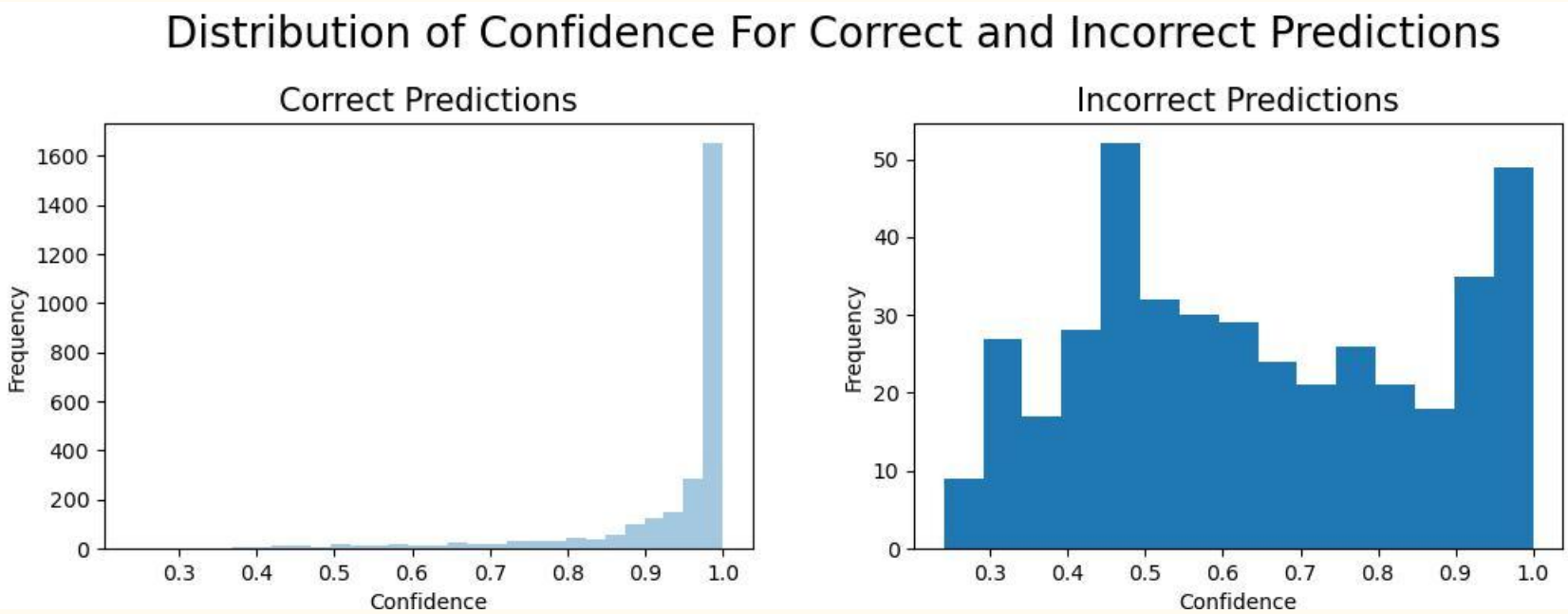
The MFCCs Convolutional Neural Network!

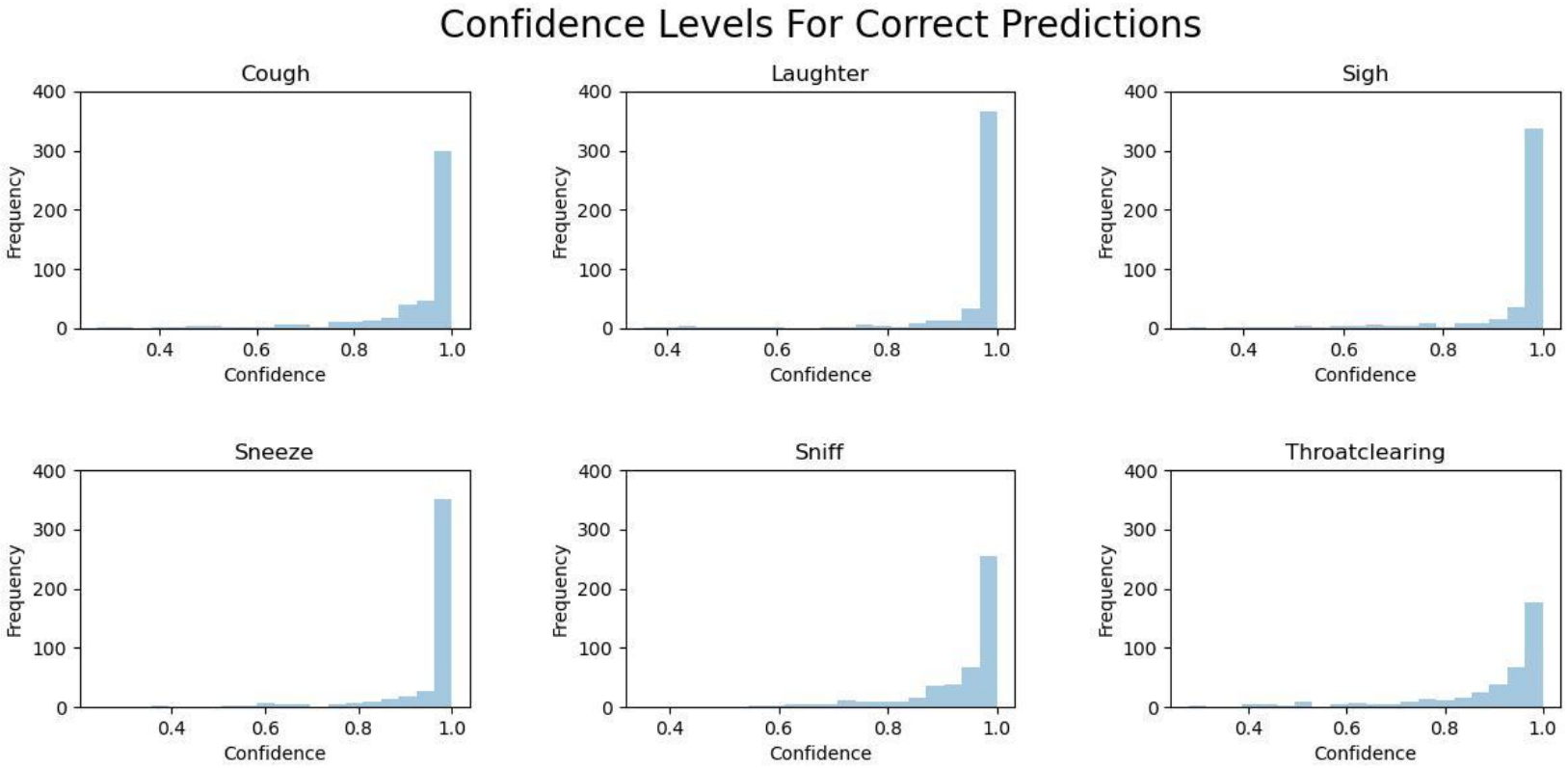
Our production model will be...

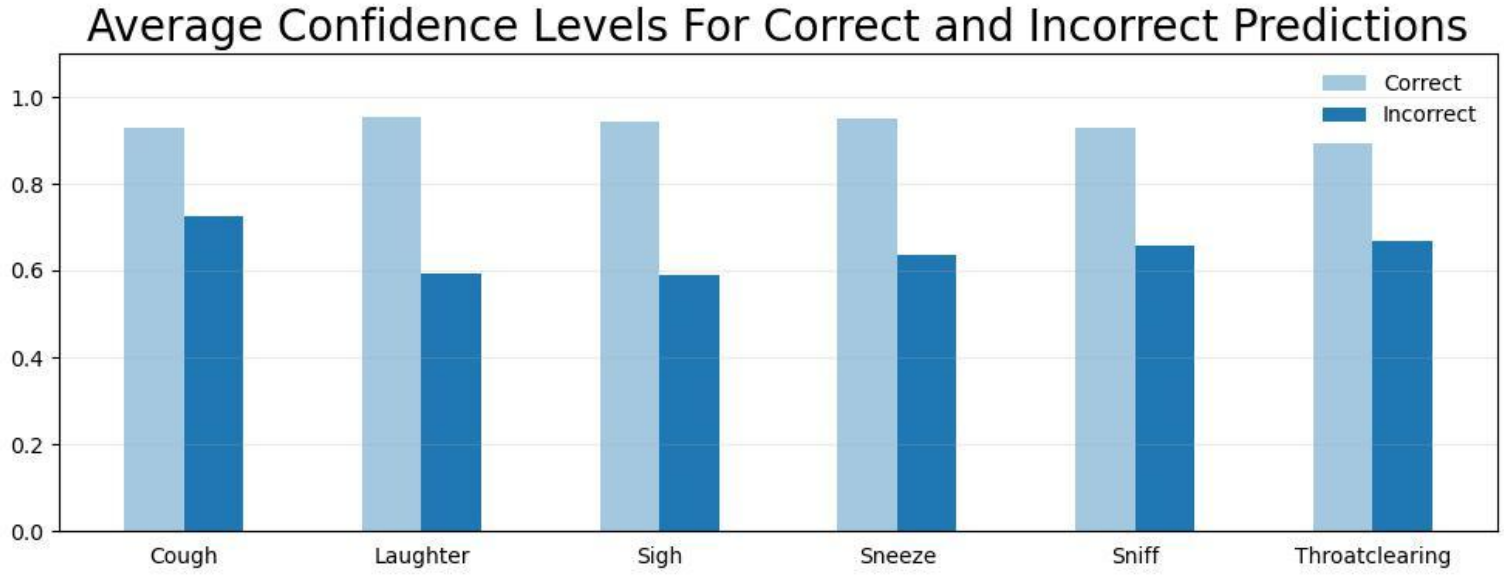
The MFCCs Convolutional Neural Network! (the first one)


Gender Does Not Influence Prediction
0.75
0.50
0.25
0.00
Correct Prediction
Incorrect Prediction
Male
Female
Other


Distribution of Ages For Correct and Incorrect Predictions
Correct Predictions
Frequency
Age
0
50
100
150
200
250
20
30
40
50
60
70
Incorrect Predictions
Frequency
Age
0
10
20
30
40
50
60
70
80
20
30
40
50
60
70


Distribution of Confidence For Correct and Incorrect Predictions
Correct Predictions
Frequency
0
200
400
600
800
1000
1200
1400
1600
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Confidence
Incorrect Predictions
Frequency
0
10
20
30
40
50
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Confidence


Confidence Levels For Correct Predictions
Cough
Laughter
Sigh
Sneeze
Sniff
Throatclearing
Frequency
Confidence
0
100
200
300
400
0.4
0.6
0.8
1.0


Average Confidence Levels For Correct and Incorrect Predictions
Correct
Incorrect
1.0
0.8
0.6
0.4
0.2
0.0
Cough
Laughter
Sigh
Sneeze
Sniff
Throatclearing

Conclusions

- We built a strong model but did not meet our goal of 90% accuracy (~87%)
- Developed a Streamlit App to share the model
- Identified where the model had the most trouble (coughs and throat clears)

Next Steps

- Collect data on other sounds to diversify the classification

- Collect more (properly labeled) audio samples for coughs and throat clears
  - Add this into Streamlit

- Look into applying on longer clips through segmenting data or sliding window
  - Strengthen/support speech recognition

To Streamlit!

Image provided by Vecteezy