AUDIO CLASSIFICATION OF EMOTIONS

Nathan McKinney,
Data Scientist,
General Assembly

INTRODUCTION

Can we accurately classify emotions given existing audio to improve student communication skills prior to technical interviews?

Can we develop a web app that will allow for predictions on new data?

7:38:55 COMMUNICATION RULE

Communication is only:

7 % verbal
93 % non-verbal.

The non-verbal component?

Body Language (55%)
Tone of Voice (38%)

FURTHER BACKGROUND

USE CASES

NEGOTIATIONS

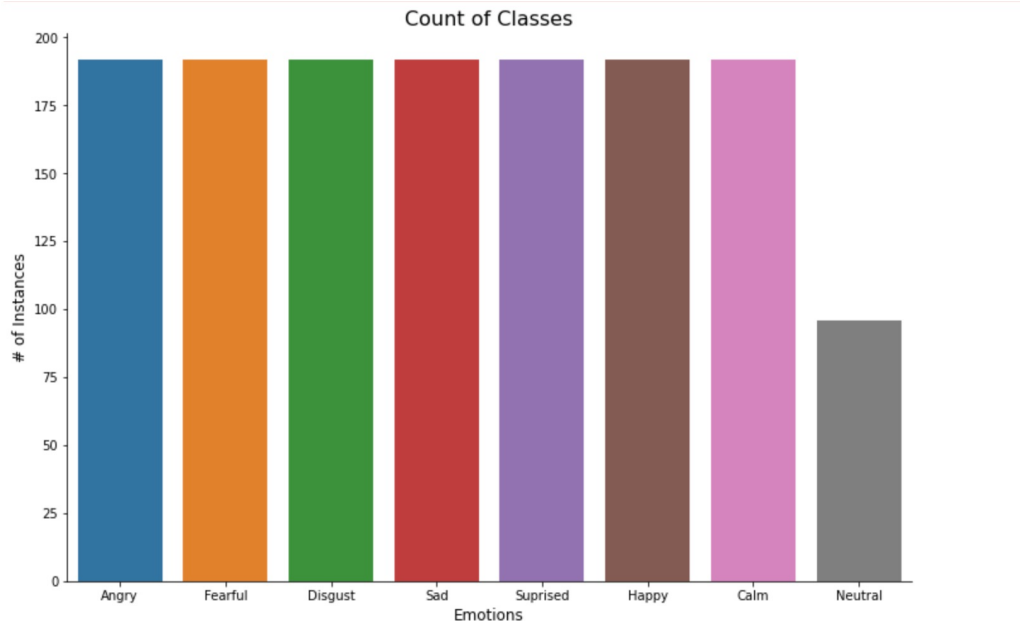
OUR SIGNAL DATA

- RAVDESS data
  - 24 Actors
  - 1440 Audio Files

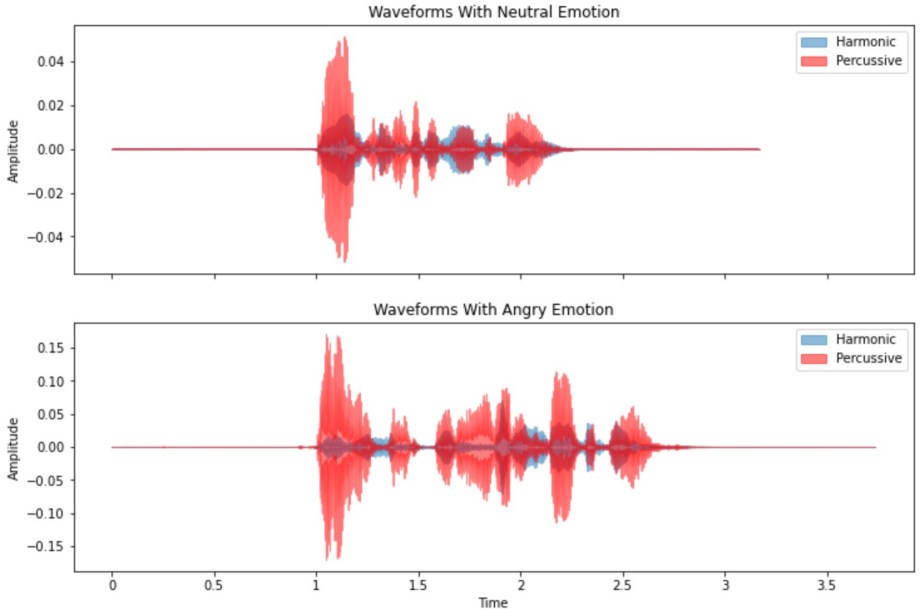
- 8 Primary Emotions
  - Neutral
  - Calm
  - Surprised
  - Happy
  - Disgust
  - Sad
  - Fearful
  - Angry

ABOUT OUR SAMPLES

8 classes of similar sample size



On average 1.5 seconds long



MODELING PROCESS

Null Hypothesis

- 12% Accuracy (Probability of us guessing an emotion correctly)
- Likely higher

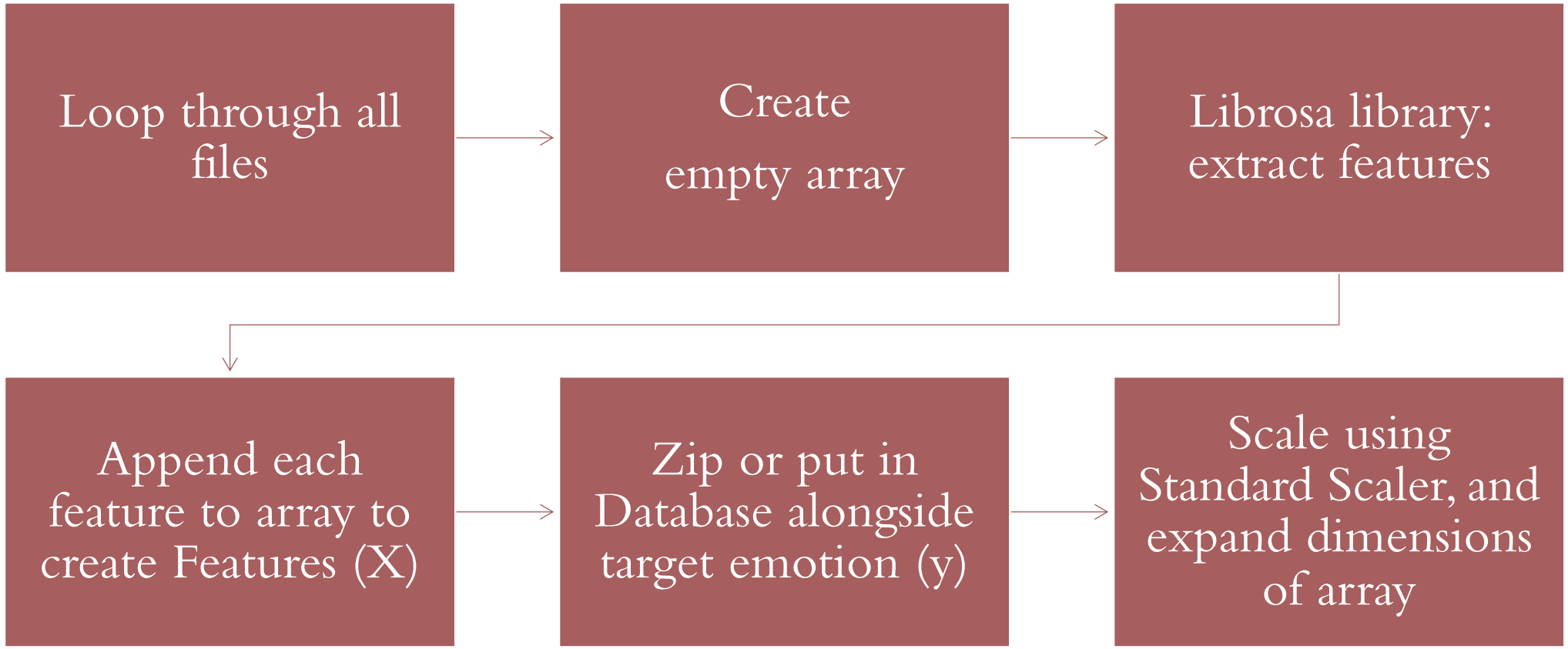
2 Different Feature Sets

- Baseline off MFCC (Explained Later)
- Multiple Features

5 Different Models

- Support Vector Machines (SVM)
- K-Nearest Neighbors
- Random Forest
- Decision Tree Classifier
- Convolutional Neural Net (CNN)

PRE-PROCESSING AUDIO DATA



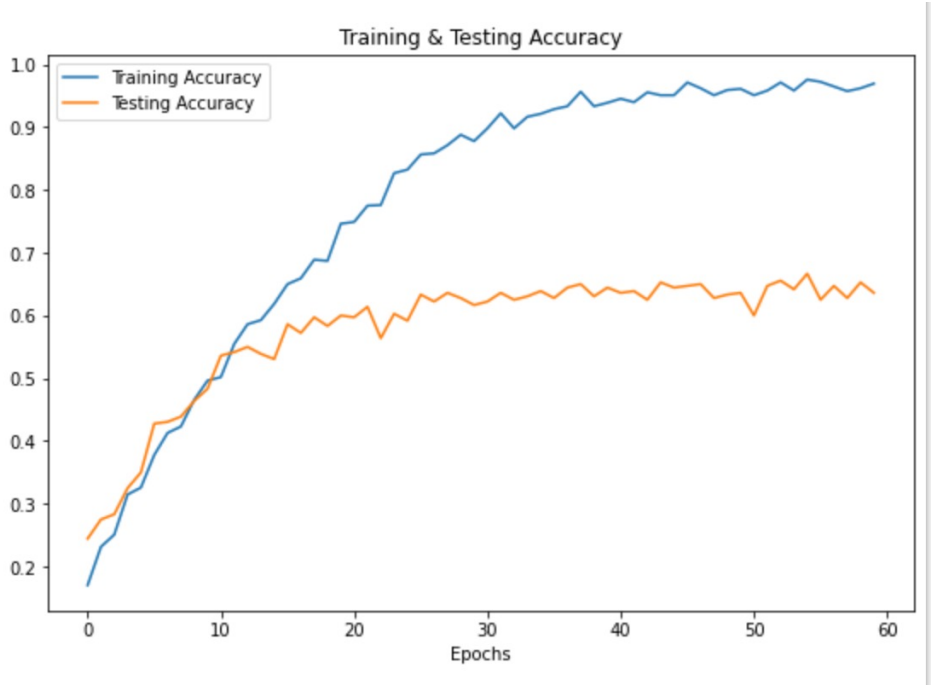
FEATURES IN SIGNAL PROCESSING

MFCC (Mel-Frequency Cepstral Coefficients):	Most commonly used; represents phonemes (distinct units of sound) coming off of the vocal tract
MEL Scale	The MEL scale is a mathematical scale that relates the perceived frequency of a tone to the actual measured frequency.
Chroma	Measurement of pitch, differentiating between 12 classes of pitch
Zero-Crossing Rate	Key feature used to classify percussive sounds

MODEL PERFORMANCE (MFCC ONLY)

- Basic framework with only one feature
- RMS prop optimizer added 7% to overall CNN scoring

Top 3	*CNN*	*SVM*	*K-Neighbors*
(Test) Accuracy	68%	63%	57%
Precision	67%	60%	56%
Recall	67%	59%	55%
F1 Score	67%	58%	54%



MODEL PERFORMANCE (MFCC + OTHER)

- Added additional features Chroma, Mel Spectrogram, etc.

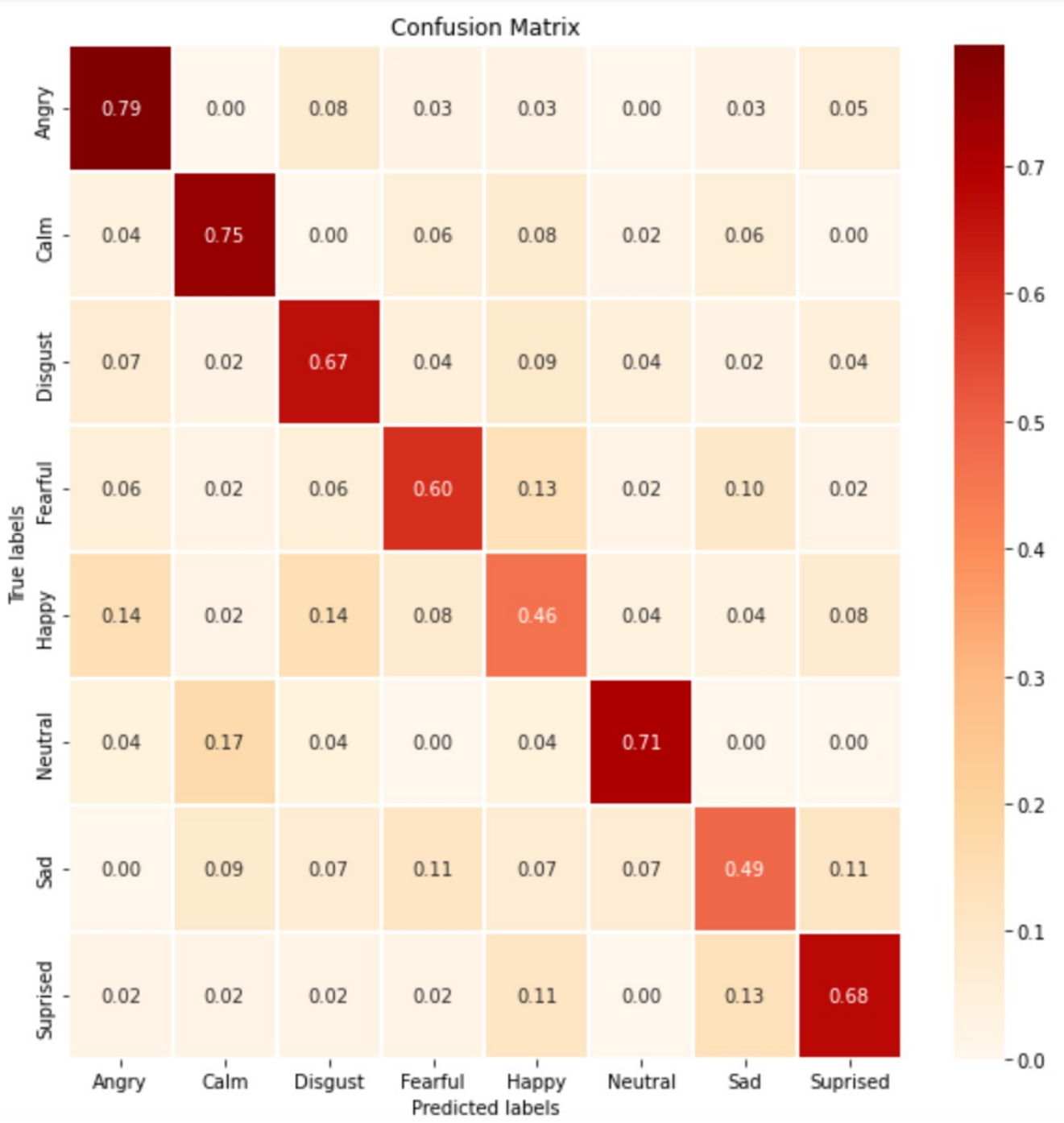
Top 3	*CNN*	*K-Neighbors*	*SVM*
(Test) Accuracy	59%	56%	52%
Precision	56%	55%	48%
Recall	57%	55%	49%
F1 Score	56%	54%	47%

	precision	recall	f1-score	support
Angry	0.74	0.52	0.61	48
Fearful	0.61	0.86	0.72	51
Disgust	0.50	0.62	0.56	48
Sad	0.49	0.43	0.45	47
Suprised	0.54	0.45	0.49	49
Happy	0.50	0.31	0.38	26
Calm	0.50	0.56	0.53	41
Neutral	0.64	0.64	0.64	50
accuracy			0.57	360
macro avg	0.56	0.55	0.55	360
weighted avg	0.57	0.57	0.56	360

ADDITIONAL FINDINGS

- Some emotions are better predicted than others
- The model had difficulty predicting happiness and sadness
- Anger and disgust were consistently mistaken for happiness

**Note: Graphic shows prediction accuracy of various features in diagonal boxes, while false predictions make up the rest*



DEMONSTRATION

CONCLUSIONS

- Our MFCC-only CNN model trains above our null hypothesis given a randomized guess, but with a caveat that will require additional improvement
- The model is not yet deployment ready
- Sometimes less is more, as using the most common baseline metric (MFCC) yielded higher performance than incorporating more features.
- Our ability to predict the tones of happiness and sadness is limited in comparison to other features, and are key identifiers we need to be able to classify moving forward.

NEXT STEPS

Improve

Improve model scoring using further data augmentation and optimization

- Pitch, add white noise, stretch

Add

Improve overfitting by adding additional data to train on

- SAVEE, TESS, CREMA-D : Sources provide an additional 10,800 audio samples

Deploy

Further develop app to be able to predict emotions of audio uploads in real time