# Visual and Lingual Emotion Recognition using Deep Learning Techniques

By

Akshay Kajale

Advisor : Dr. Chris Pollett

Committee : Dr. Robert Chun
Mr. Kiran Salte

# Agenda

- Introduction
- Background
- Convolutional Neural Network
- Depthwise Convolutional Neural Network
- MFCCs
- Chroma Features
- Mel Spectrogram
- System Architecture
    - Facial Emotion Recognition
    - Speech Emotion Recognition
    - Fusion Algorithm
- Android Application
- Dataset
- Class Activation Maps
- Experiments and Results
- Conclusion and future work

# Introduction

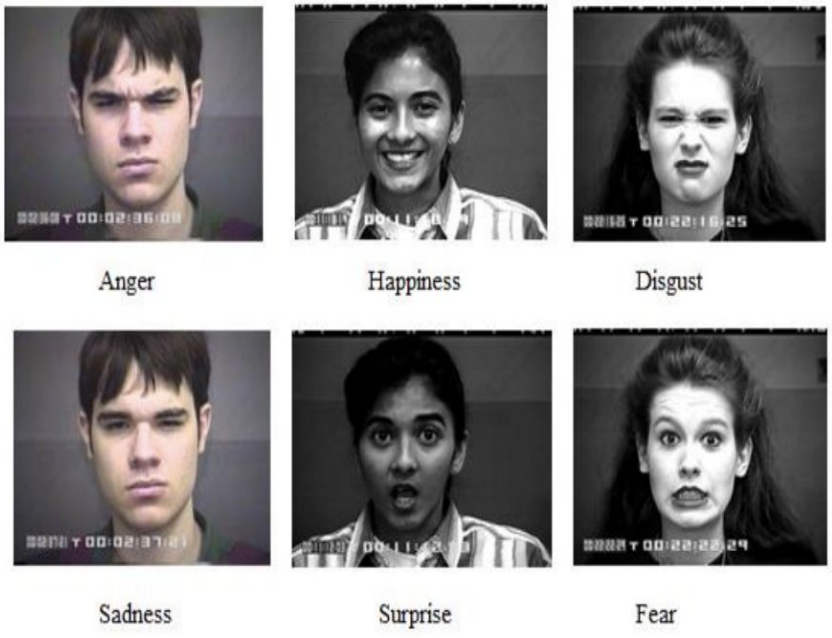
- Human Emotion can be recognized by various attributes like Facial Expressions, Hand Gestures, Pitch of the Speech, Text.
- The objective of this project is to build an Android application to identify the human emotion based on facial expressions and speech tone.
- Application can access both cameras at the same time and also able to record the conversation.
- This allows application to predict the emotion of the overall conversation.
- Our system classifies emotions into seven classes: Neutral, Surprise, Happy, Fear, Sad, Disgust, and Angry.

## Background

- Six emotions are universally experienced in humans: Angry, Happy, Disgust, Sad, Surprise and Fear.
- Emotions can be expressed verbally (words and tone of voice) or non-verbally (Body language and facial expressions).
- Charles Darwin believed that humans and animals show emotion through remarkably same behavior.

# Background (Cont'd)

- Emotions can be expressed practicing various face attributes.
- Happy: The eyebrows are relaxed, mouth is open and mouth corners are upturned.
- Sad: The inner eyebrows are bent upward, eyes are slightly closed, mouth is relaxed.
- Facial expressions can be deceiving, Fear can be confused for Surprise.



Anger

Happiness

Disgust

Sadness

Surprise

Fear

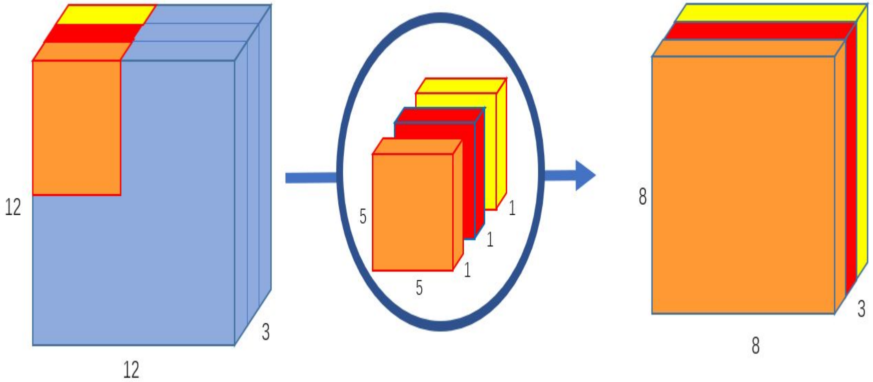
*The six primary emotional states (Kanade, Cohn, and Tian, 2000)*

# Convolutional Neural Network

- Takes Image as an Input.
- Assigns Importance (Learnable weights and Biases) to various aspects of the Image.
- Preprocessing required is much lower as compared to other classification models.
- 2D Convolution performed over multiple input channels.
- Filter is as deep as the input and lets us freely mix channels to generate each element in the output.

# Depthwise Convolutional Neural Network

- Apply Single convolutional filter for each input channel.
- Depthwise convolution keeps each kernel separate.
  - Split the Input and Filter into channels.
  - Convolve each input with the respective filter.
  - Stack the convolved outputs together.



*Chi-Feng Wang ,Depthwise Convolution*

# Background (Speech)

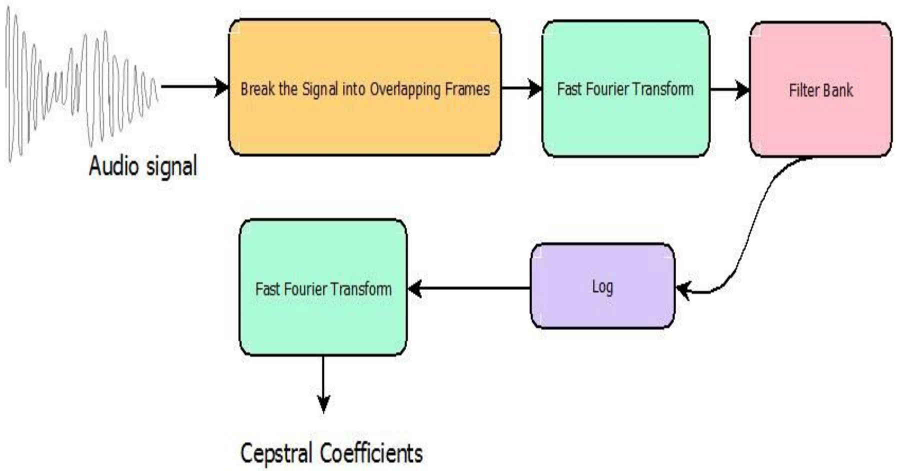
- Speech is one of the  most decisive indicators for emotion recognition.
- There are numerous features in human speech which are used to identify emotion.
- Mel Frequency Cepstral Coefficient (MFCC), Chroma Feature, and Mel Spectrogram.

*All Answers Ltd. (November 2018). Model of Speech Recognition Using MFCC Extraction. Retrieved from https://ukdiss.com/examples/speech-recognition-using-mfcc.php?vref=1*

*https://medium.com/analytics-vidhya/understanding-the-mel-spectrogram-fca2afa2ce53#:~:text=You%20can%20think%20of%20a,over%20time%20at%20different%20frequencies.*

# MFCCs

- Coefficients which represents audio based on perception.



*https://medium.com/prathena/the-dummys-guide-to-mfcc-aceab2450fd*

# Chroma Features

- Chroma Feature is a descriptor which represents the tonal content of the audio signal in condensed form.
- Primary Property is that they occupy the harmonic and melodic characteristic of voice irrespective of changes in timbre and instrumentation.

*Shah, Ayush & Kattel, Manasi & Nepal, Araju & Shrestha, D.. (2019). Chroma Feature Extraction.*

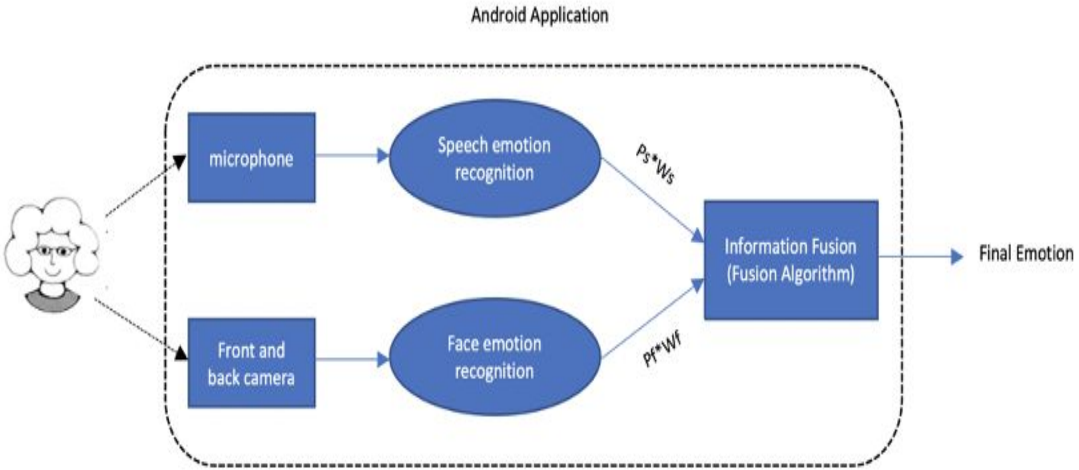
# Mel Spectrogram (Cont’d)

- Visual representations of FFT is called as Spectrogram.
- Studies have shown that Humans do not perceive frequencies on Linear Scale.
- We can easily tell the difference between 500 and 1000 Hz but hardly be able to tell the difference between 10,000 and 10,500 Hz even though distance between the pairs is same.
- Stevens, Volkmann, and Newmann proposed a unit of pitch such that equal distances in pitch sounded equally distant to the listener, called as Mel Scale.
- The Spectrogram represented in Mel Scale is called as Mel Spectrogram.

*Leland Roberts, https://medium.com/analytics-vidhya/understanding-the-mel-spectrogram-fca2afa2ce53*

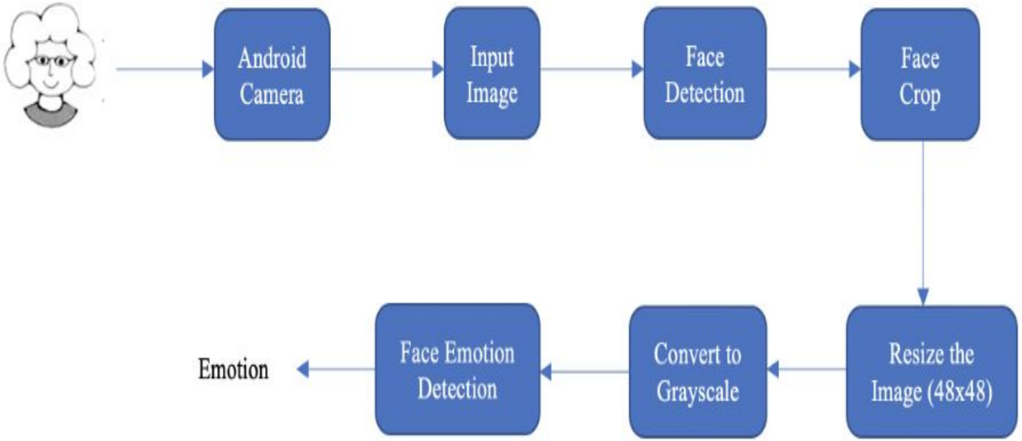
# Architecture

- The Architecture is divided into three Segments
  - Facial Emotion Recognition
  - Speech Emotion Recognition
  - Information Fusion



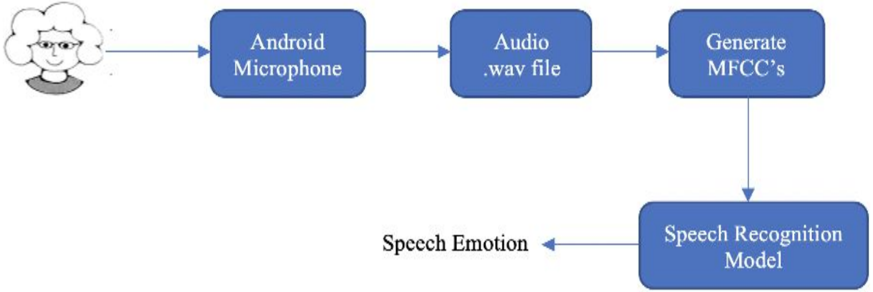
# Facial Emotion Recognition

- Front and Back Camera captures one image/frame sequentially.
- Faces are captured and stored from those images using Viola.
- Stores images are reshaped into 48 X 48.
- Face images are normalized and converted to grayscale.
- These images are passed to tflite model in form of TensorImage.
- The model returns the array of emotions with probability.



# Speech Emotion Recognition

- Application records the conversation between two people for every 5 seconds.
- The recording is stored in WAV format.
- The recorded file is used to generate MFCCs.
- MFCCs are passed to speech emotion recognition model.
- The model returns the array of emotions with probability.



*Microsoft Corporation (June 1998). "WAVE and AVI Codec Registries - RFC 2361". IETF. Retrieved 2009-12-06.*

# Fusion Algorithm

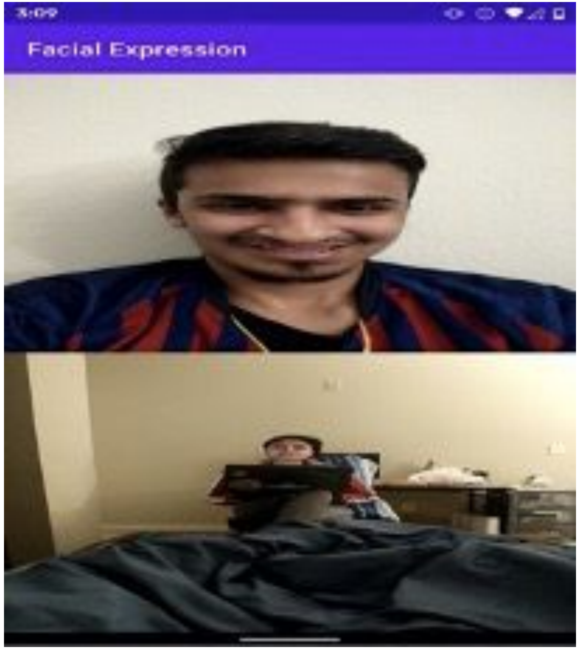
- Used to associate emotions obtained from facial and speech model and predict one emotion for a period.
- Count the number of emotions given by facial and speech model over the period.
- Store the topmost emotion probability and the respective emotion given by both models. For example, $PF_{max}$ = 7/10 and $F_{emo}$ = Happy. Similarly let us consider $PS_{max}$ = 5/10 and $S_{emo}$ = Neutral.
- 55% of emotions and attitudes are represented by facial expression; speech shows 38% and 7% is related to the spoken words.
- Consider constant weights $W_f = 0.55$ and $W_s = 0.38$.
- Assign the $Emo_{Final}$ as max $(W_f * PF_{max,}\ W_s * PS_{max})$.

*C. Marechal et al., « Survey on AI-Based Multimodal Methods for Emotion Detection », in High-Performance Modelling and Simulation for Big Data Applications: Selected Results of the COST Action IC1406 cHiPSet, J. Kołodziej et H. González-Vélez, Éd. Cham: Springer International Publishing, 2019, p. 307‑324.*

*Yao, Qingmei, "Multi-Sensory Emotion Recognition with Speech and Facial Expression" (2014). Dissertations. 710.*
*https://aquila.usm.edu/dissertations/710*

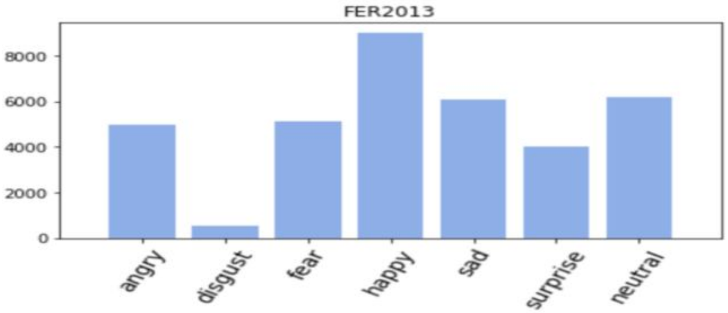
# Android Application

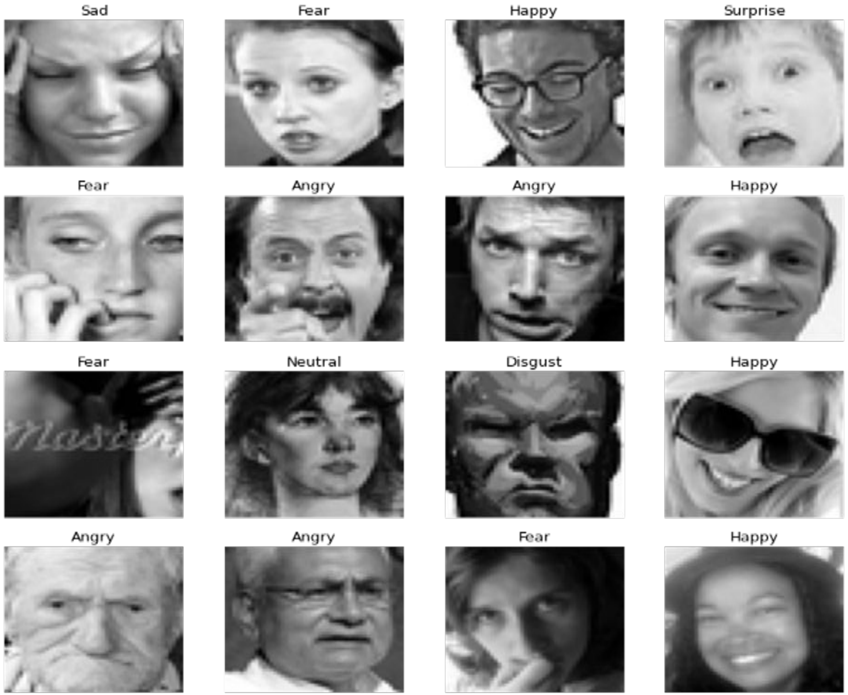
- The application is developed in Java.
- Both models are developed in Keras and trained on Google Colab.
- Keras models are converted into tflite models due to limited computing power of mobile.
- These lite models are deployed on Android application to get desired result.



# Dataset: Facial Expressions

- FER-2013 Dataset, contains 35,587 Images of people displaying various emotions through facial expressions.
- Training Images: 28,709, Testing Images: 7178.

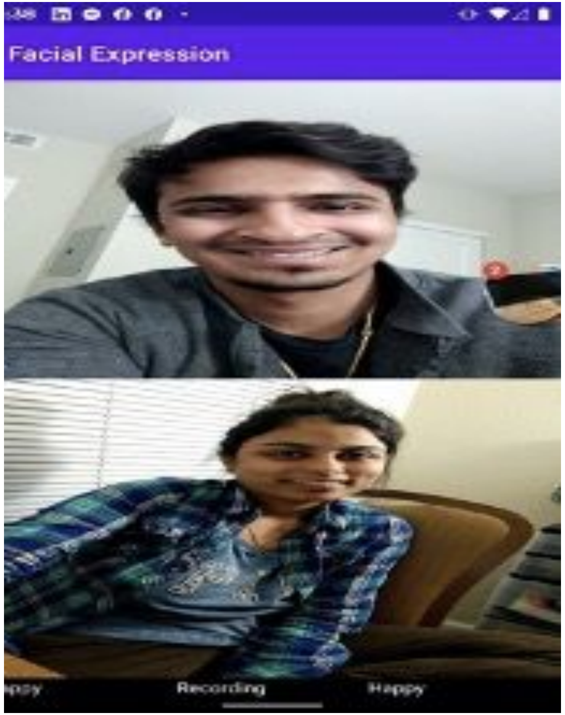




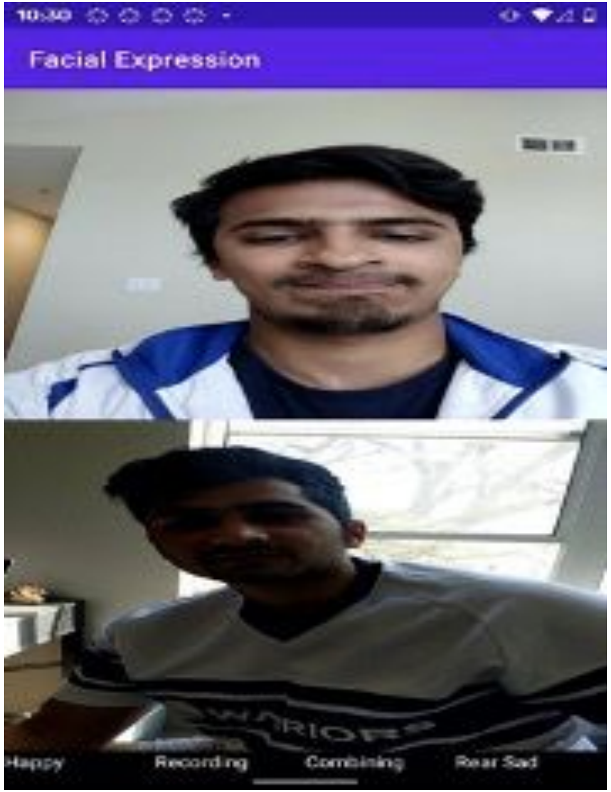
# Dataset: Speech

- Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS).
- 605 audio recordings expressing different emotions using a speech by 24 different actors.
- Training Samples: 453, Testing Samples: 152.
- The Audio format is WAV.
- Samples are classified into four classes: Happy, Sad, Neutral and Angry.

# Experiments

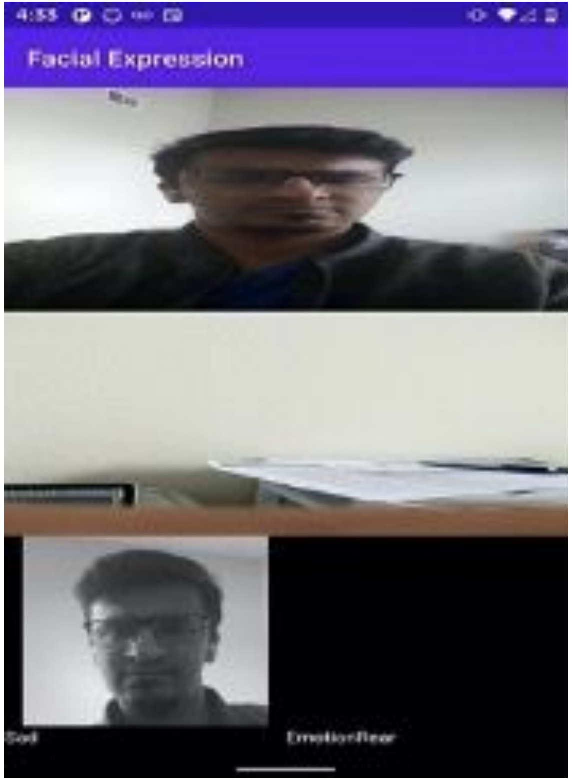


Front: Happy, Recoding, Rear: Happy

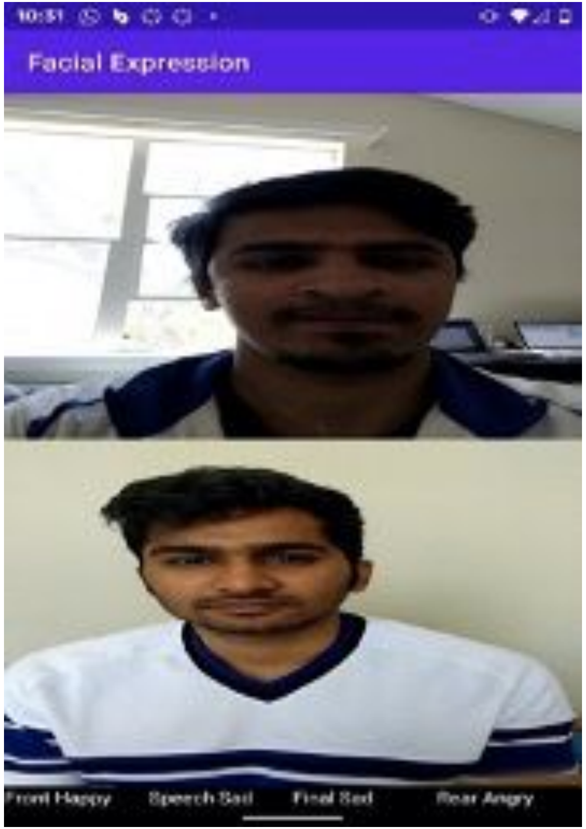


Front: Happy, Recording, Combining, Rear Sad

# Experiments: (Cont'd)

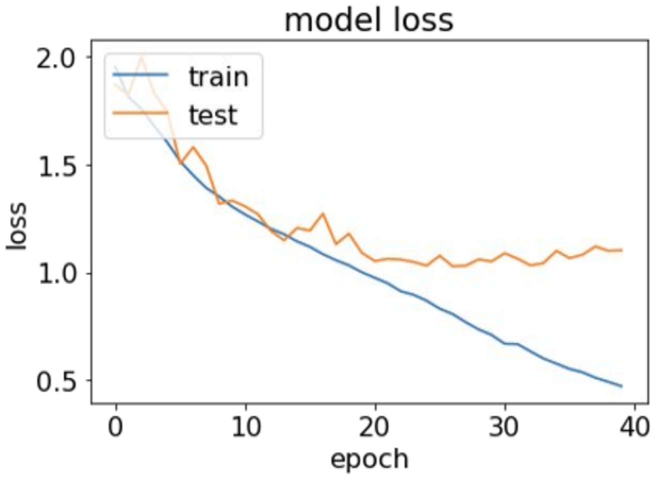


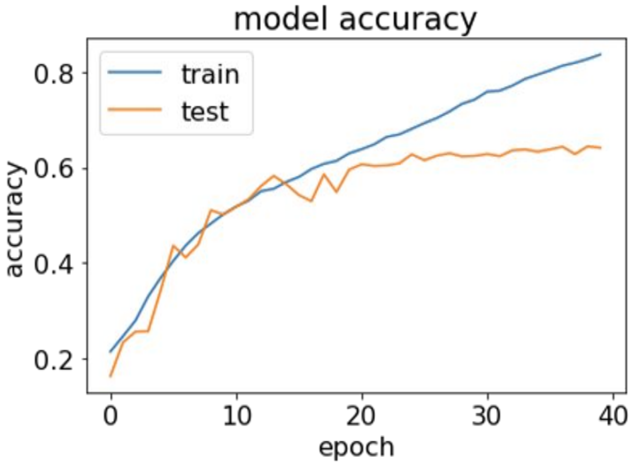
Sad, EmotionRear

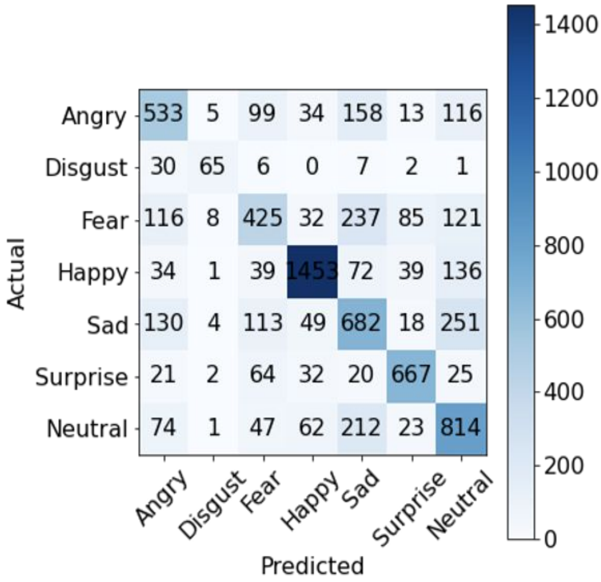


Front: Happy, Speech: Sad, Final:Sad, Rear:Angry

# Results: Face Emotion Recognition (Sequential Model)





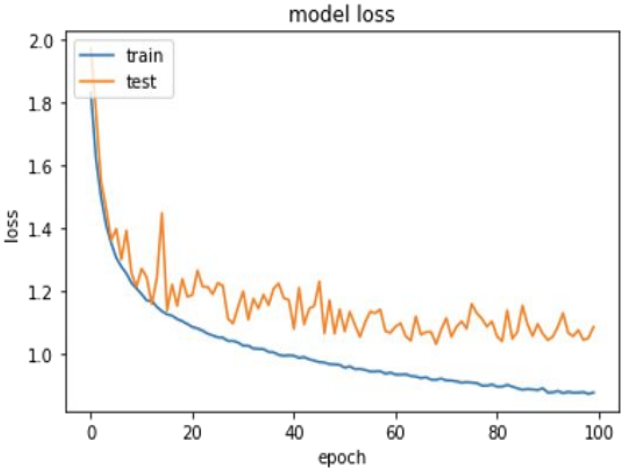


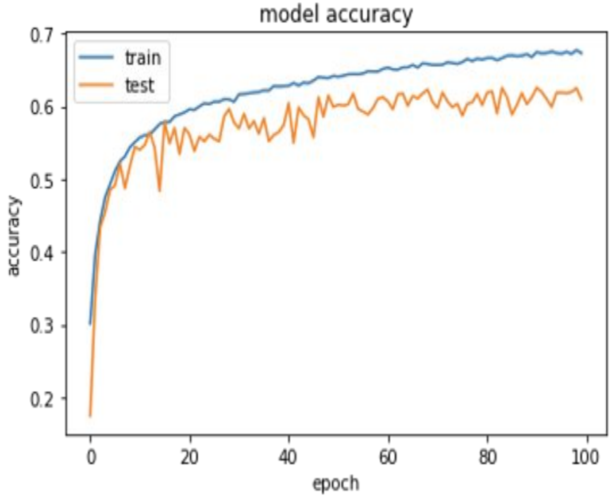
| Actual \ Predicted | Angry | Disgust | Fear | Happy | Sad | Surprise | Neutral |
|---|---|---|---|---|---|---|---|
| Angry | 533 | 5 | 99 | 34 | 158 | 13 | 116 |
| Disgust | 30 | 65 | 6 | 0 | 7 | 2 | 1 |
| Fear | 116 | 8 | 425 | 32 | 237 | 85 | 121 |
| Happy | 34 | 1 | 39 | 1453 | 72 | 39 | 136 |
| Sad | 130 | 4 | 113 | 49 | 682 | 18 | 251 |
| Surprise | 21 | 2 | 64 | 32 | 20 | 667 | 25 |
| Neutral | 74 | 1 | 47 | 62 | 212 | 23 | 814 |

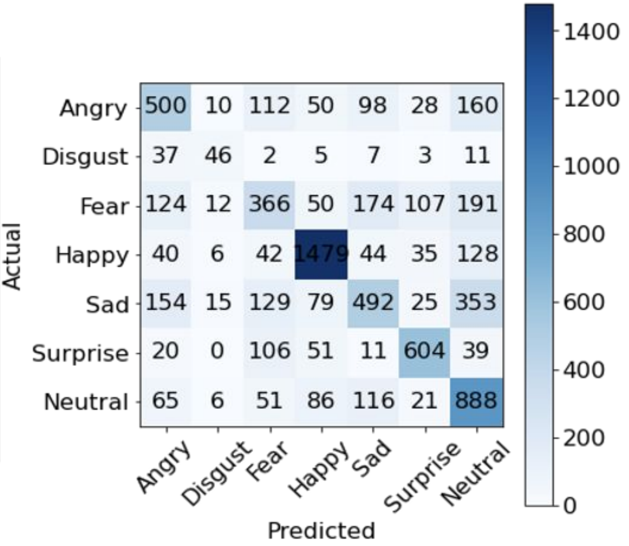
## Results: Face Emotion Recognition (Sequential Model)

- Testing Accuracy: 64.62%.
- Best Classified Emotion: Happy, accuracy 81.90%.
- Worst Classified Emotion: Fear, accuracy 41.50%.
- Fear is misclassified into Sad and Neutral.

# Results: Face Emotion Recognition (DepthWise Model)







| | Angry | Disgust | Fear | Happy | Sad | Surprise | Neutral |
|---|---|---|---|---|---|---|---|
| Angry | 500 | 10 | 112 | 50 | 98 | 28 | 160 |
| Disgust | 37 | 46 | 2 | 5 | 7 | 3 | 11 |
| Fear | 124 | 12 | 366 | 50 | 174 | 107 | 191 |
| Happy | 40 | 6 | 42 | 1479 | 44 | 35 | 128 |
| Sad | 154 | 15 | 129 | 79 | 492 | 25 | 353 |
| Surprise | 20 | 0 | 106 | 51 | 11 | 604 | 39 |
| Neutral | 65 | 6 | 51 | 86 | 116 | 21 | 888 |

## Results: Face Emotion Recognition (DepthWise Model)

- Testing Accuracy: 60.95%.
- Best Classified Emotion: Happy, accuracy 83.37%.
- Worst Classified Emotion: Fear, accuracy 35.74%.
- Fear is misclassified into Sad and Neutral.

# Depth wise vs Sequential Model

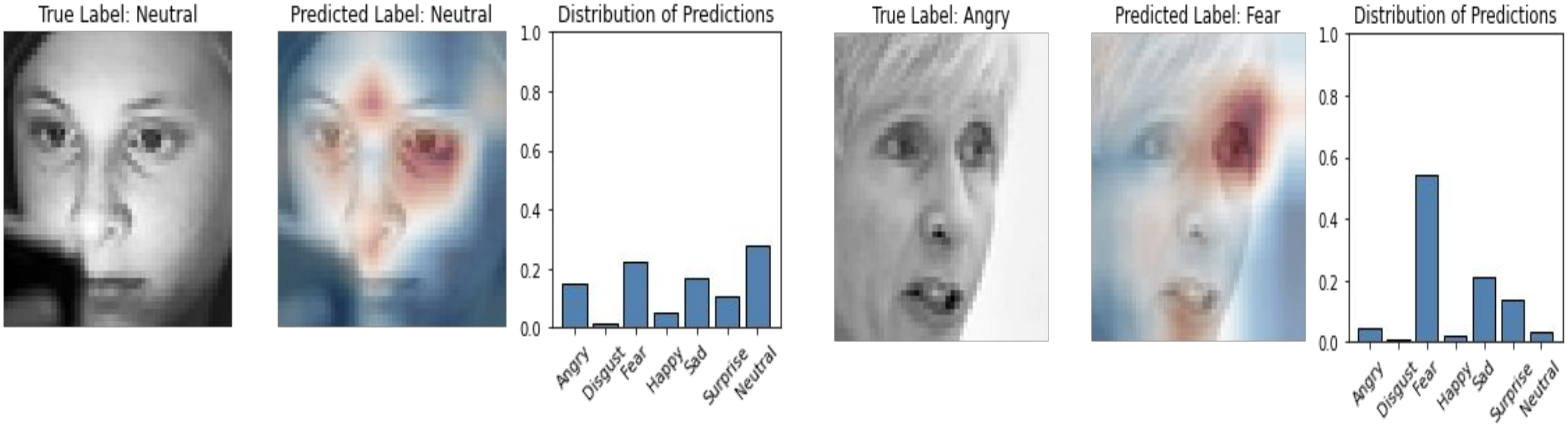
| Model | Training Accuracy | Testing Accuracy |
|---|---|---|
| Sequential | 84.22% | 64.62% |
| Depthwise Convolution | 67.75% | 60.95% |

# Why Depth-Wise Over Sequential Model?

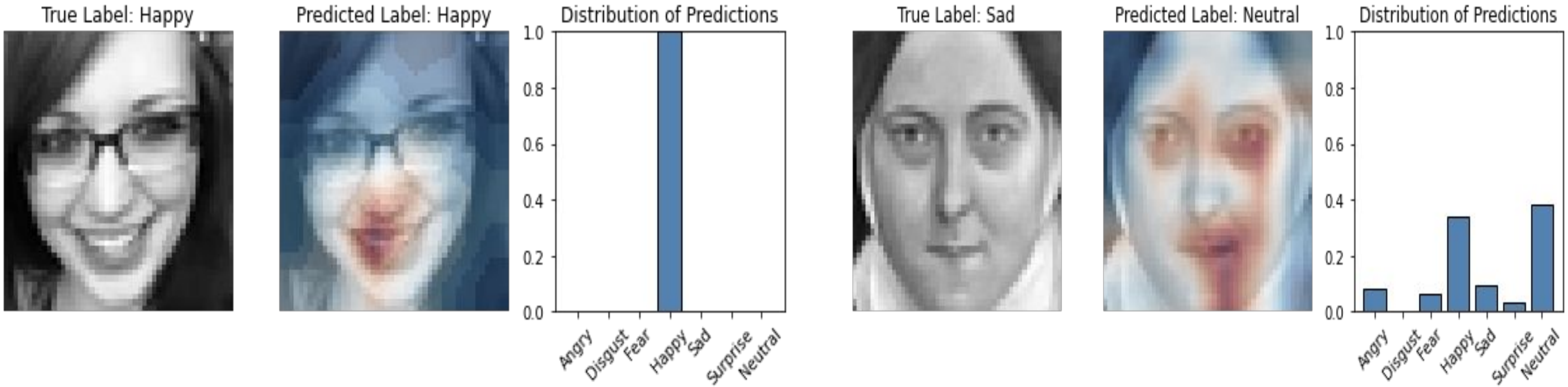
- Sequential Multiplications / Depth Wise Multiplications = $(1/L + 1/D_K{}^2)$ .
- L = Number of filters, $D_K$ = Kernel Size.
- Consider L = 1024, and $D_K$ = 3, the ratio obtained is 1/9.
- Normal Convolution requires 9 times more multiplications than Depth-Wise Convolution.
- Since we are deploying our model on mobile which has limited resources, less number of multiplications is better for the performance of the application.
- Application gave performance issue (Crash, delay in producing output, delay in importing model) when Sequential model was used.
- Hence we choose Depth-Wise model to deploy on Android application.

*CodeEmporium, Depth Wise Separable Convolution- A FASTER CONVOLUTION* *https://www.youtube.com/watch?v=T7o3xvJLuHk*

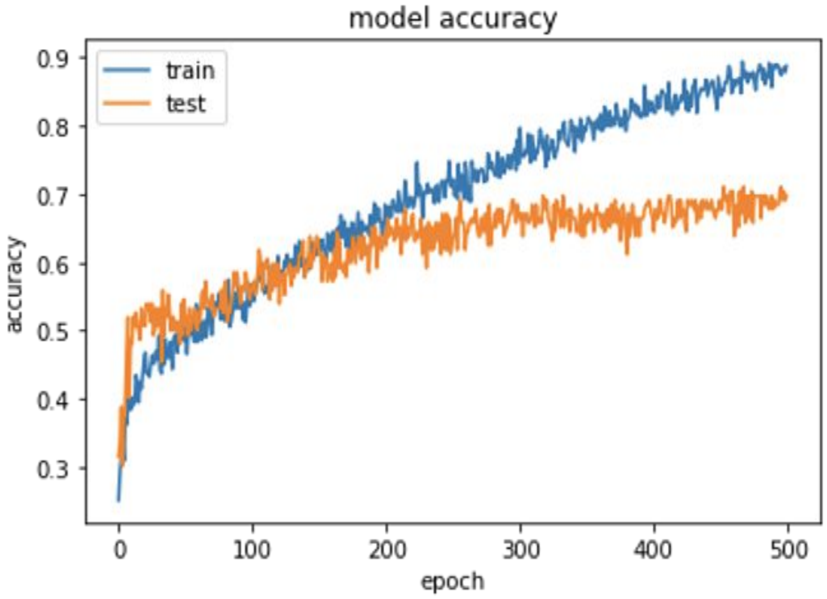
# Class Activation Map

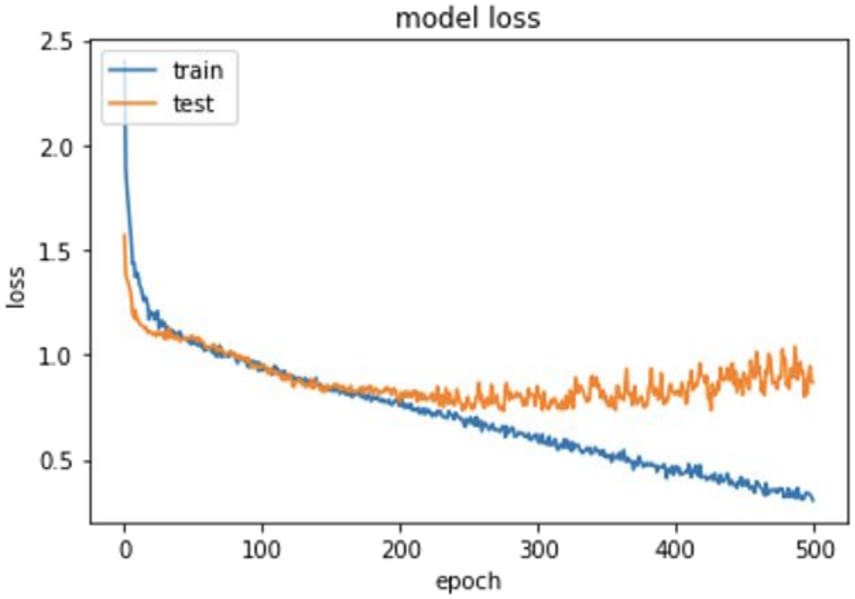


# Class Activation Map (Cont'd)

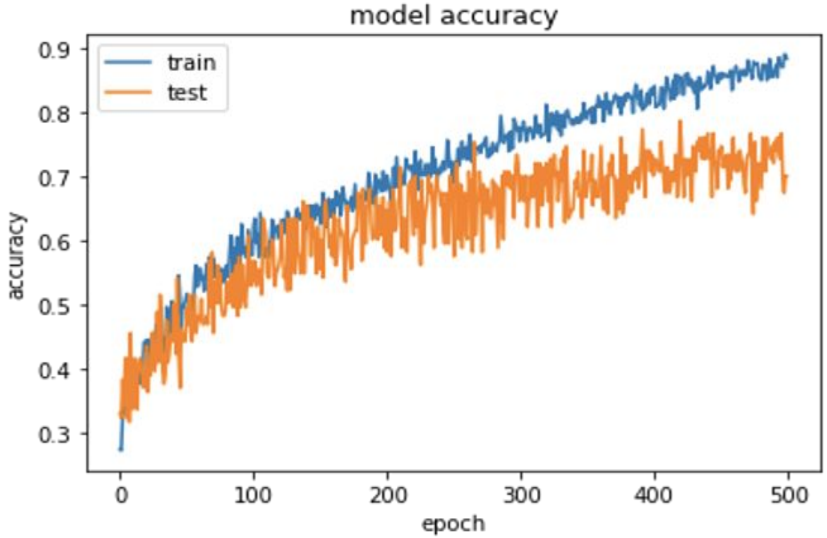


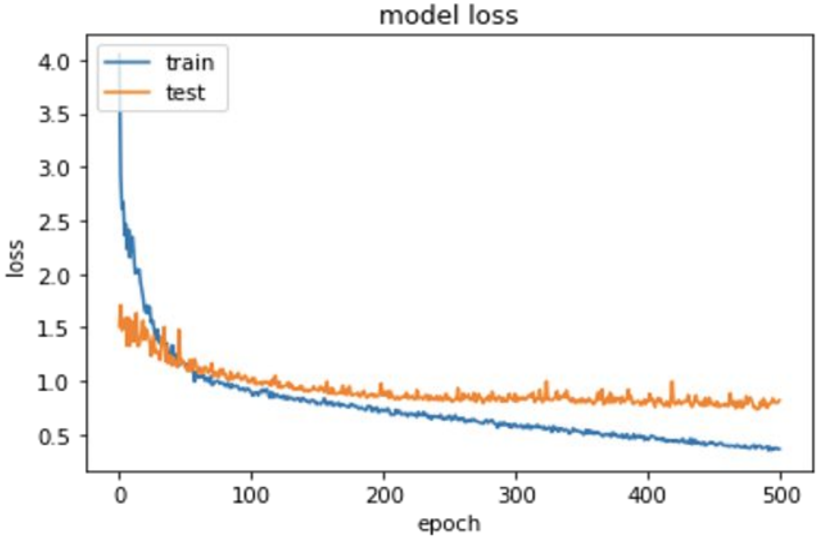
# Results: Speech Emotion Recognition (MFCCs, Chroma and Mel Spectrogram)



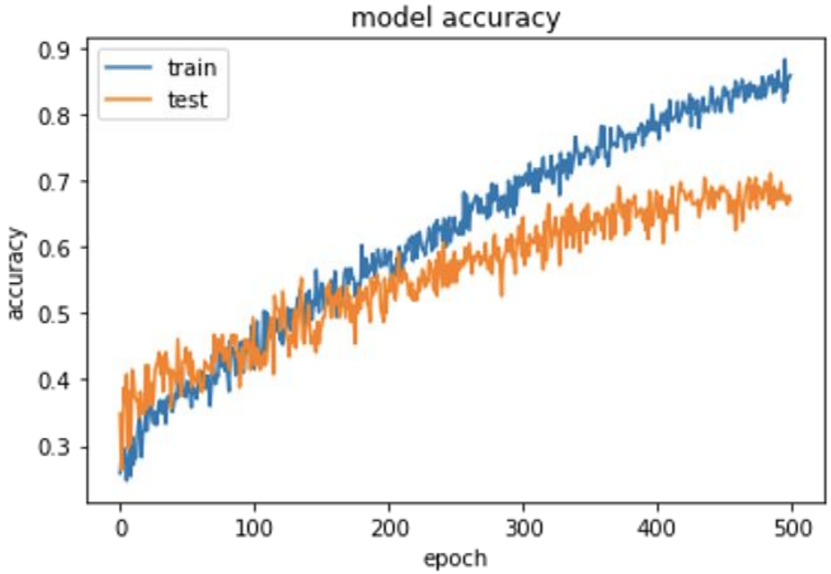


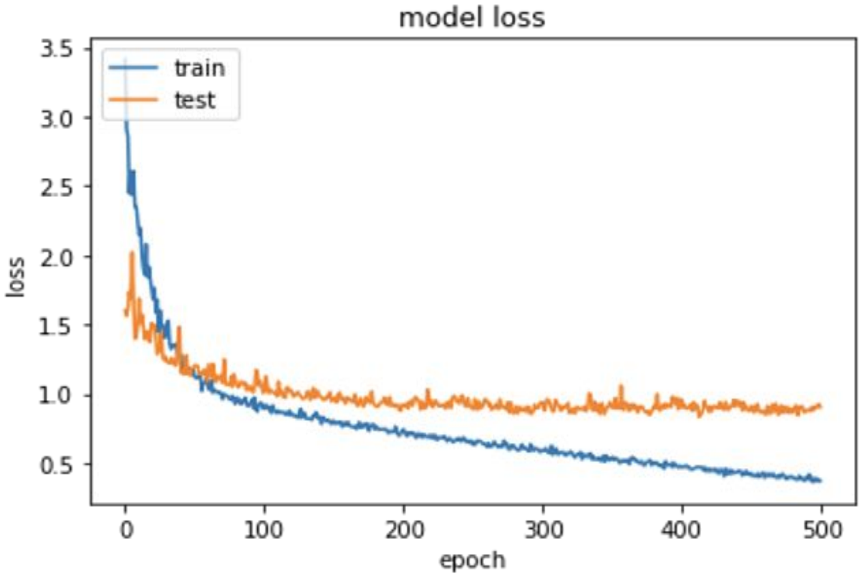
# Results: Speech Emotion Recognition (MFCCs and Mel Spectrogram)





# Results: Speech Emotion Recognition (MFCCs)

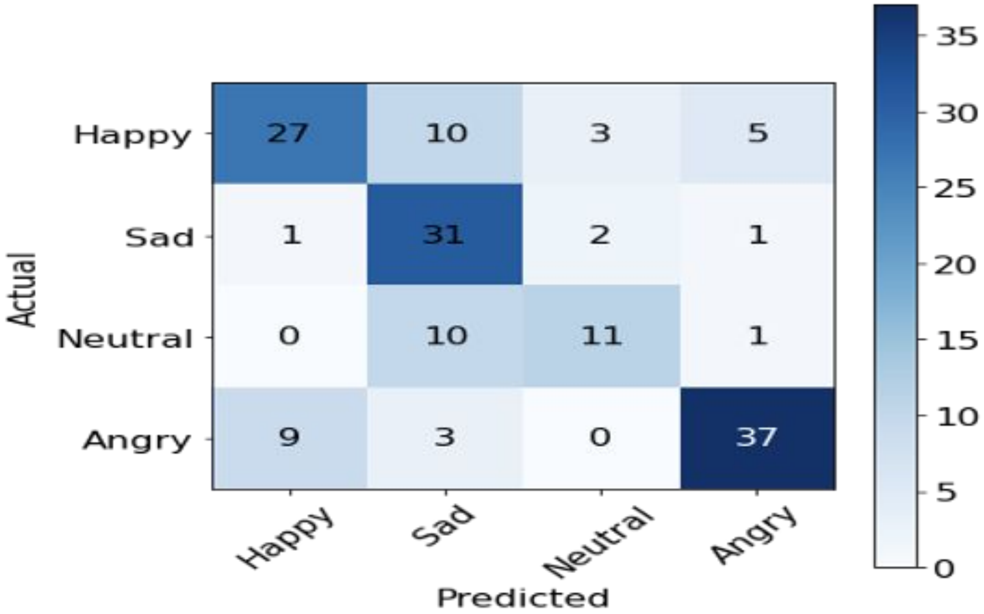




# Accuracy Comparison Based On Features

| Features | Training Accuracy | Testing Accuracy |
|---|---|---|
| MFCCs, Chroma and Mel Spectrogram | 89.82% | 69.62% |
| MFCCs and Mel Spectrogram | 88.54% | 68.67% |
| MFCCs | 87.95% | 67.52% |

# Confusion Matrix (MFCCs)



| Actual \ Predicted | Happy | Sad | Neutral | Angry |
|---|---|---|---|---|
| Happy | 27 | 10 | 3 | 5 |
| Sad | 1 | 31 | 2 | 1 |
| Neutral | 0 | 10 | 11 | 1 |
| Angry | 9 | 3 | 0 | 37 |

## Why we used only MFCCs to Predict Emotion Based on Speech?

- There is no significant difference between accuracies.
- Python uses Librosa library to calculate MFCCs, Chroma features and Mel Spectrogram.
- For Android, there is a substitute library named JLibrosa, which provides similar features.
- The library is still under development. It does not provides the functionality to calculate Chroma features and Mel Spectrograms.

# Conclusion

- Multimodal emotion recognition looks promising as compared to single model.
- Depth wise Model performs better as compared to Sequential Model in terms of resource utilization.
- Classic Machine Learning Techniques did not perform well on Audio Dataset.
- Facial attributes has more weightage than any other attribute when predicting emotion.
- Tflite models perform better as compared to Keras models when there is a resource constraint.

# Future Work

- More emotion data can be aggregated from a greater number of people to train and test the facial and speech models to improve their accuracy.
- Include more Human Attributes in the emotion prediction. For example, Hand Gestures, Body Gestures, Speech Text etc.
- The UI and performance of the application can be improved.

# Demo

“A Demo is worth 1000 Pictures”

# References

1. All Answers Ltd. (November 2018). Model of Speech Recognition Using MFCC Extraction. Retrieved from https://ukdiss.com/examples/speech-recognition-using-mfcc.php?vref=1

2. “Understanding the Mel Spectogram” https://medium.com/analytics-vidhya/understanding-the-mel-spectrogram-fca2afa2ce53#:~:text=You%20can%20think%20of%20a,over%20time%20at%20different%20frequencies

3. CodeEmporium, Depth Wise Separable Convolution- A FASTER CONVOLUTION https://www.youtube.com/watch?v=T7o3xvJLuHk

4. Microsoft Corporation (June 1998). "WAVE and AVI Codec Registries - RFC 2361". IETF. Retrieved 2009-12-06.

5. C. Marechal et al., « Survey on AI-Based Multimodal Methods for Emotion Detection », in High-Performance Modelling and Simulation for Big Data Applications: Selected Results of the COST Action IC1406 cHiPSet, J. Kołodziej et H. González-Vélez, Éd. Cham: Springer International Publishing, 2019, p. 307-324.

6. Yao, Qingmei, "Multi-Sensory Emotion Recognition with Speech and Facial Expression" (2014). Dissertations. 710. https://aquila.usm.edu/dissertations/710