Spring 2021

Visual and Lingual Emotion Recognition using Deep Learning Techniques

Akshay Kajale
San Jose State University

Visual and Lingual Emotion Recognition using Deep Learning Techniques

A Project Report

Presented to

Christopher Pollett

Robert Chun

Kiran Salte

Department of Computer Science

San José State University

In Partial Fulfillment

Of the Requirements for the class

CS298

By

Akshay Kajale

May 2021

The Designated Project Committee Approves the Master's Project Titled

Visual and Lingual Emotion Recognition Using Deep Learning Techniques

By

Akshay Kajale

APPROVED FOR THE DEPARTMENT OF COMPUTER SCIENCE

SAN JOSE STATE UNIVERSITY

MAY 2021

Dr. Christopher Pollett	Department of Computer Science
Dr Robert Chun	Department of Computer Science
Mr. Kiran Salte	Jabil Inc

Acknowledgment

I would like to thank my advisor, Dr. Chris Pollett, for all his help and guidance, and motivation that he has given me over the past year. I consider myself extremely fortunate to have had an opportunity to work with someone as humble and learned as him.

I would like to express my gratitude to the members of my defense committee, Dr. Robert Chun and Kiran Salte. I would also like to thank the CS faculty for supporting me during these two years of my graduate study

Finally, I would like to thank my parents and friends for all the support and motivation over the years.

Abstract

Emotion recognition has been an integral part of many applications like video games, cognitive computing, and human computer interaction. Emotion can be recognized by many sources including speech, facial expressions, hand gestures and textual attributes. We have developed a prototype emotion recognition system using computer vision and natural language processing techniques. Our goal hybrid system uses mobile camera frames and features abstracted from speech named Mel Frequency Cepstral Coefficient (MFCC) to recognize the emotion of a person. To acknowledge the emotions based on facial expressions, we have developed a Convolutional Neural Network (CNN) model, which has an accuracy of 68%. To recognize emotions based on Speech MFCCs, we have developed a sequential model with an accuracy of 69%. Out Android application can access the front and back camera simultaneously. This allows our application to predict the emotion of the overall conversation happening between the people facing both cameras. The application is also able to record the audio conversation between those people. The two emotions predicted (Face and Speech) are merged into one single emotion using the Fusion Algorithm. Our models are converted to TensorFlow-lite models to reduce the model size and support the limited processing power of mobile. Our system classifies emotions into seven classes: neutral, surprise, happy, fear, sad, disgust, and angry

Index Terms: **Deep Learning, Depth-wise convolutional model, Mel Frequency Cepstral Coefficient, Sequential Model**

List of Figures and Tables

Table of Contents

I. Introduction

Emotions command our day-to-day activities; they are a big part of the human experience, communication, and decision-making. For example, humans tend to rehash the activities that make them feel happy and comfortable but skip on the sad activities. Emotion is a comprehensive psychological and physiological process related to many factors such as emotion, temperament, personality, and motivation [1]. This report describes a project to develop an Android application that can detect emotions based on facial and speech features. This application can access both cameras simultaneously. This functionality will allow us to understand the overall emotion of conversation between the person facing the front and back camera. This application is also able to record the conversation. Audio will be used to predict the emotion based on speech features. The emotions obtained (Speech and Facial) are then merged into single emotion using the Fusion Algorithm.

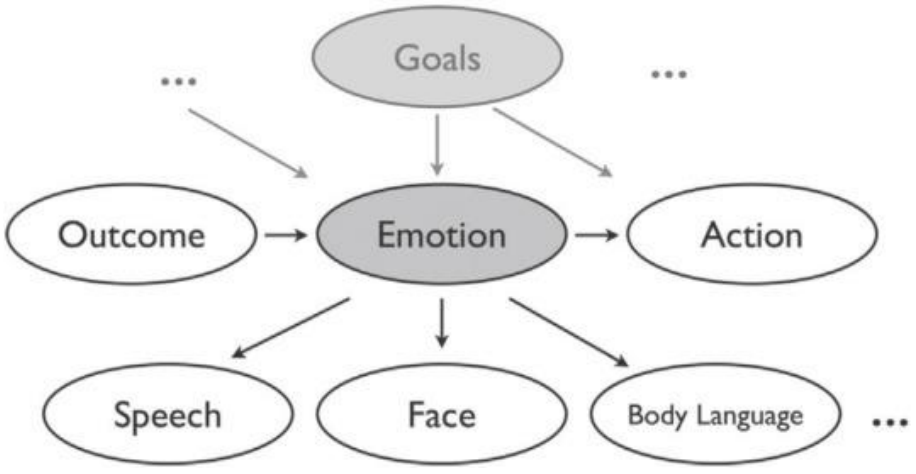

Figure 1. Model of lay theory (Ong et al.,2015)

Emotions can be recognized by many methods like vital signals, gesture recognition, empathy, etc. The most effective solution is to recognize emotion by two types of features, facial

and lingual. The human face is the most exposed part of the body in daily activities. Based on the facial expressions (eyebrows, eyes, mouth, chicks), other people can anticipate how the person is feeling or trying to express. But sometimes, facial expressions can be deceiving. Hence there is a necessity for more noticeable features for emotion recognition. Speech is one of the effective ways to express emotions. Based on the selection of words and pitch of the voice, humans can recognize another person's emotion.

In the early work, Mehrabian [2] showed that 55% of emotional information is visual, 38% vocal, and 7% verbal. Facial expression change in the communication is the significant sign which indicates a change in the emotional state. Humans tend to focus on different face areas when they try to interpret what each expression might mean. For example, eyes can determine if someone is sad or angry. Open mouth indicates the sign of fear, raised corner can determine happy emotion. Corners of the mouth that are drawn down conveys Sadness.

As speech is the easiest and effective form of communication for human beings, in this technologically advanced world, it is natural that we would expect our machines to understand us based on our speech. Still, apart from the message, the speech carries secondary level information also such as gender, emotional states [3]. An important issue in speech emotion recognition is the extraction of efficient speech features that characterize the emotional content of speech [4]. Both facial and speech features have their advantages and disadvantages. To solve this problem, we aim to develop an Android Application that predicts a person's emotion based on both facial and lingual features. The output of both facial and speech emotion models will be provided to the Fusion Algorithm as an input to find the final emotion.

Now we discuss the organization of this report. In the following chapters, we will look at the background of emotion recognition and the techniques used to detect emotions based on facial

and lingual features. Chapter 2 contains the background of emotion recognition. Chapter 3 will see the architecture design and process flow of the Android application developed for emotion recognition. In Chapter 4, we will see the deep learning models to recognize emotions based on face and speech. Chapter 5 describes the results obtained from the application. Chapter 6 describes the conclusion and future work.

II. Background

According to psychologist Paul Eckman, there is a total of six emotions that are universally experienced in all human cultures [5]. Those are identified as Happy, Sad, Disgust, Fear, Surprise, and Anger. Fig. 2 shows facial expressions representing basic emotions.

Figure 2. The six primary emotional states (Kanade, Cohn, and Tian, 2000)

Emotion can be expressed majorly in two ways.

2.1 Facial Expression

Every emotion has its way of expressing using facial features. There are many techniques based on which emotions can be detected. E. Ghaleb, M. Popa, and S. Asteriadis [6] have proposed Multimodal Emotion Recognition Metric Learning (MERMEL) for the challenging audio-video emotion recognition task. They have used RBF-based SVM for the classification of emotions-based face and speech modalities and achieved the accuracy of 65.2%. D. Kalita [7] has used

geometric feature-based method (geometric location of eyes, eyebrow, nose lips, etc.) to classify the emotions. They have categorized their emotions into six primary emotional states.

Charles Darwin was one of the first to study human and animal emotions. In his early work, he published a book named “Expression of the Emotions in Man and Animals," He argued that humans and animals show emotion through remarkably similar behavior [16]. Darwin was beyond any doubt that internal feelings of humans were outwardly manifested similarly to animals. He stated an example where he observed that angry emotion drives to eye contraction and teeth exposure in humans as well as animals. He believed that such expressions must have developed through common progressive mechanisms. Darwin was inclined towards use of photograph since he believed photographs will have an advantage over other representation shapes since its potential for capturing transitory expressions with precision and separation would justify objective [16]. Darwin's theory can be strong corroboration to the assumption that this Android application can identify animal emotions.

To understand this project, we must understand the concepts of facial expressions and speech pitch. For facial emotion recognition, the main part of understanding is how the model classifies the emotions into multiple classes. In the preprocessing, the image is converted into grayscale, and the model identifies the differentiating parts, which classifies the images into seven emotions. For example, in the case of happy emotion, the model will focus more mouth and teeth part of the image. In case of angry or sad emotions, the model concentrates more on eyes and lips. Similarly, for every emotion, some points differentiate them from each other. The model focuses on those differentiating points.

2.2 Speech Expressions

Speech is one of the important indicators for emotion recognition. There are multiple features or properties in the human speech based on which emotion can be identified. It has been observed in the results that as we include a greater number of features for prediction, there is an increase in the accuracy of the model.

2.2.1 Mel Frequency Cepstral Coefficient

In novice words, MFCC can be defined as a spectrum of the log of the spectrum which is called as cepstrum. Periodic factor is represented by the Fourier spectrum in a conventional analysis. This Fourier spectrum is obtained by applying Fourier transform to the speech signal [20]. The next step is to take the log of the magnitude of this Fourier spectrum. The last step is to take a spectrum of this log by a cosine transformation. The obtained spectrum does not lie within the frequency domain or time domain [20]. Hence, Bogert, et al. decided to call it the quefrency domain. This obtained spectrum was named Mel Frequency Cepstral Coefficient.

2.2.2 Chroma Feature

The chroma feature is a descriptor representing the tonal content of a musical audio signal in a condensed form. Short-Time Fourier Transform (STFT) and Constant Q Transform (CQT) are used for chroma feature extraction [14]. STFT is used to determine frequency information of signal changes over time. CQT converts a data series to a frequency domain. The primary property of chroma feature is that they occupy the harmonic and melodic characteristic of voice irrespective of changes in timbre and instrumentation.

2.2.3 Mel Spectrogram

In Mel Spectrogram, the frequencies are transformed into Mel Scale. Studies have determined that humans do not identify frequencies on a linear scale. We are better at detecting differences in lower frequencies than higher frequencies [15]. For example, humans can easily identify differences between 100 and 500 Hz. Still, as the frequency increases, it gets hard to identify the difference even if the distance between the two pairs is the same. Based on this feature, emotions based on speech can be identified.

III. Architecture

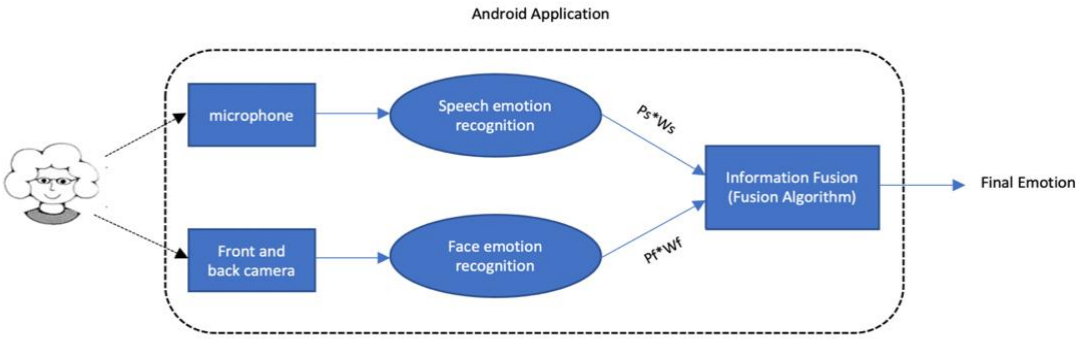

Figure 3. System Architecture

Fig. 3 shows the architecture of the Android application. Two or more users will be facing and communicating with each other. This application will be running, and one of them will be holding the mobile. The facial images are captured from both cameras simultaneously and passed to the facial emotion recognition tflite model. The facial emotion recognition model returns the emotion with the highest probability. This application also records the conversation between the people and stores it in a WAV [17] file. The MFCCs are extracted from this WAV [17] file using the JLibrosa library. These MFCCs are passed to the speech emotion recognition model. The speech model returns emotion with the highest probability. Both outputs are passed to a fusion algorithm to find the single emotion.

The overall architecture is divided into three parts:

1. Facial Emotion Recognition.
2. Speech Emotion Recognition.
3. Information Fusion.

3.1 Facial Emotion Recognition

Understanding facial expression is an imperative portion of non-verbal communication. People who are physically dumb express their emotions using hand gestures and facial expressions. Humans tend to identify how a person is feeling based on its facial expression. Mr. Bean and Charlie Chaplin are perfect examples of how emotions can be expressed using face.

Table 1: Description of Facial Expression of six basic emotions (Sumpeno et al., 2011)

No	Emotion Name	Description of Facial Expression
1	Happy	Wrinkles around the eyes. The mouth is kept open, and the lip corners are lifted diagonally.
2	Sad	The inner eyebrows are raised. The eyes are loose. Lip corners are either pulled down or remains as it is.
3	Fear	The eyebrows are raised. The mouth and eyes are wide opened.
4	Anger	Eyebrows are pulled down and together. The eyes are wide open. There is narrowing of lips corner or opened to expose the teeth.
5	Surprise	Similar signs as fear. The eyebrows are raised. The mouth and eyes are wide opened. A small hint of smile when mouth is opened.
6	Disgust	The eyes are relaxed. The upper and lower lips are curled.

Sometimes, human facial expressions can be deceiving. Du, Shichuan, and Aleix M Martinez [8], based on the experiments, have concluded that fear can be confused for surprise but not vice versa. Hence there is a need for a technique that can identify these emotions correctly. Researchers have been working on this problem for a decade. Machine learning and deep learning techniques have been used to solve this problem. S. Ramakrishnan and I. Emary [9] have used YOLO V3 Deep Neural Network on JAFEE and CK+ datasets to classify emotions. They tested their model on the JAFEE dataset and have achieved an overall accuracy of 98.12%. But the model

is very heavy to train; hence it is not suitable to deploy on devices having limited computation power. Qingyang Zhu, Guanming Lu, and Jingjie Yan [10] have used physiological data such as EEG signals, peripheral physiological signals, and facial expressions to predict the emotion. They are trying to prove that multimodal fusion achieves better results than unimodal and bimodal.

We developed two models for recognizing facial expressions, namely sequential model, and depth-wise convolutional model, to experiment and finalize the model that we will deploy on the mobile. According to Cindi May [21] Female face is more expressive than man face. Women do smile more than men, and there is evidence that women exaggerate facial expressions for positive emotions. McDuff and colleagues conducted a small experiment where they used 2000 face samples to explore different facial expressions. They used an automated facial coding system to evaluate those expressions. The results of the study confirmed that women are more expressive than men. Hence, we anticipated that there might be the requirement of two different models to recognize emotions of men and women individually. But based on dataset and machine learning concepts, we concluded that if we trained our model on large number of images having equal men and women faces, single model would be able to predict expressions of both.

3.2 Model Architecture

3.2.1 Sequential Model

The sequential model consists of one input layer, three max-pooling, and six convolutional 2D layers.

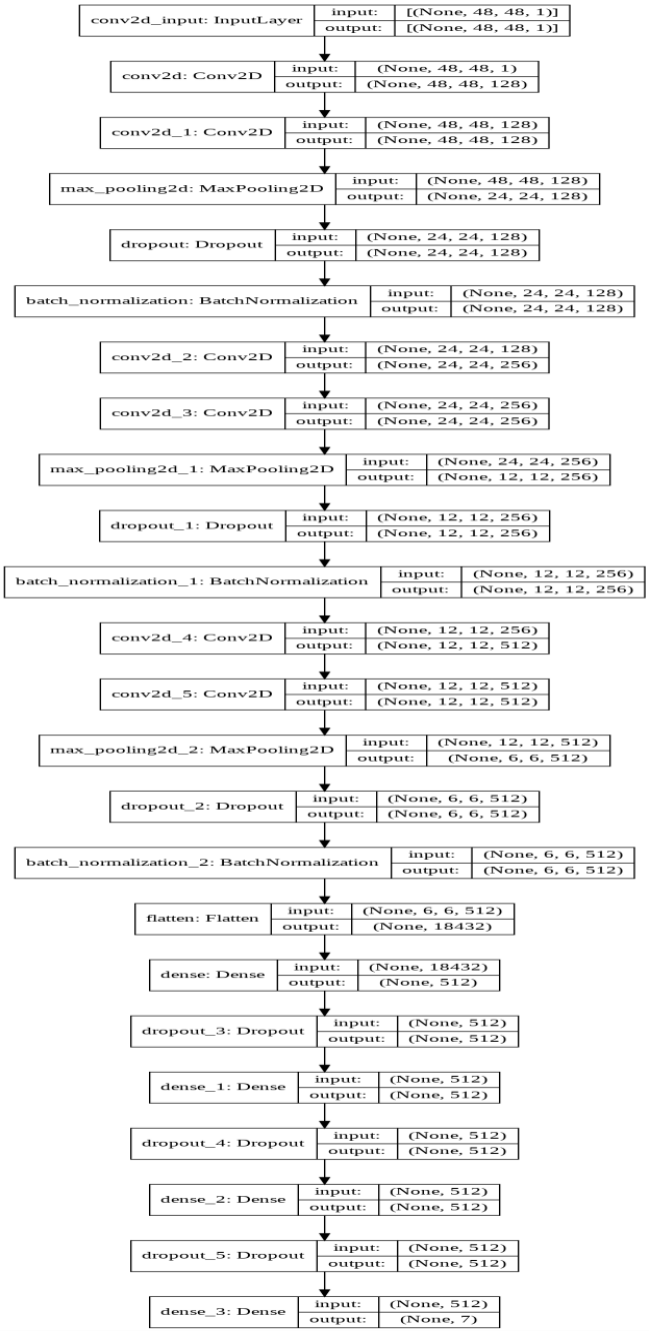

Figure 4. Sequential Model Architecture

3.2.2 Depth-wise convolutional model

Depth wise model contains two convolutional 2D layers, four Residual 2D layers.

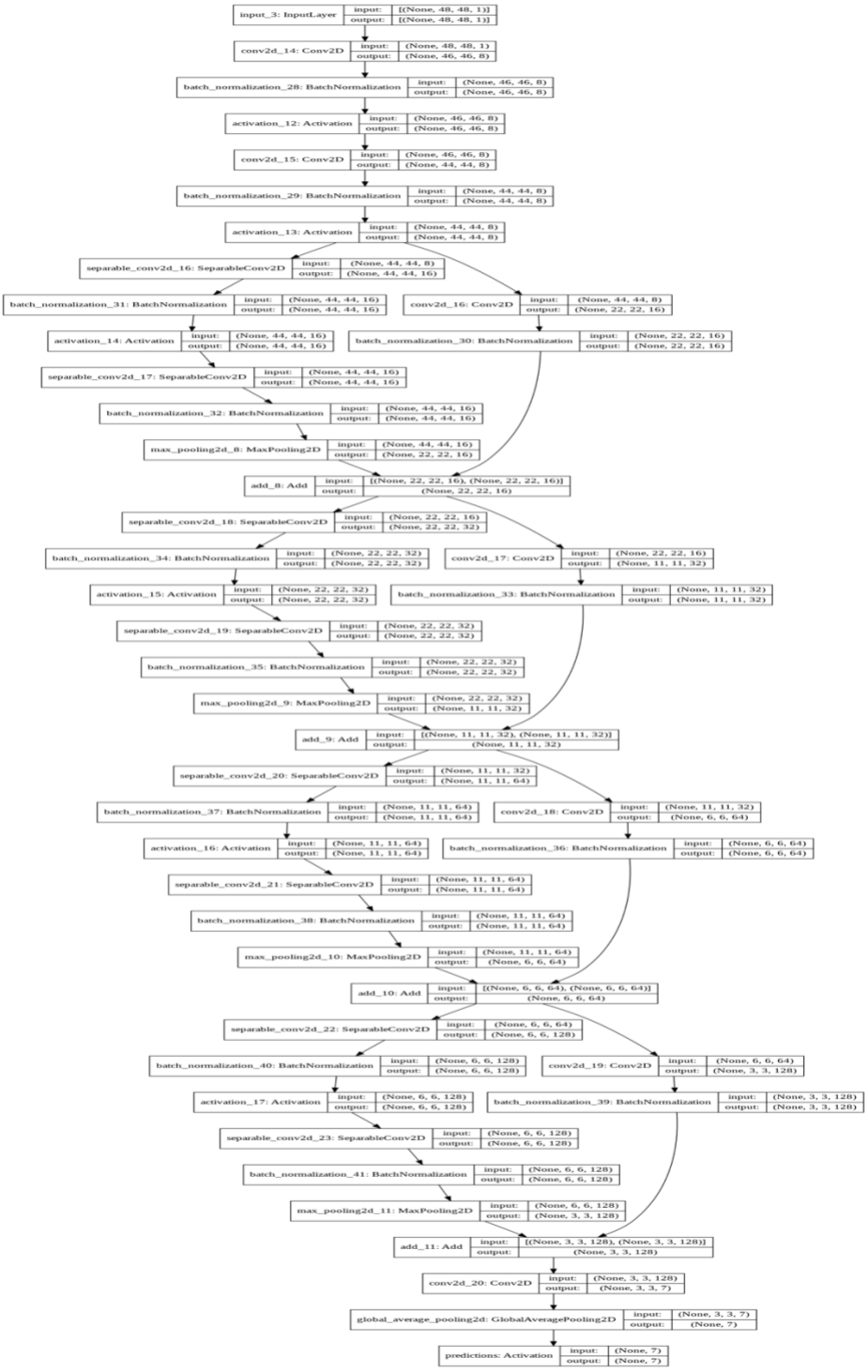

Figure 5. Depth Wise Convolutional Model Architecture

Over the years, Convolutional Neural Network (CNN) has driven enormous accomplishments, especially in deep learning [9]. Since CNN has been introduced, there has been a boom in the research of domains like computer vision and natural language processing. But CNN comes at a cost. Generally, CNN consists of multiple layers. Some of those layers are used to summarize your image to an abstract representation that can be used for classification. However, traditional CNN, like sequential, are expensive in terms of resources. This is one of the major reasons to use depth wise convolutional model. Since mobile has limited computational resources, a sequential model would have been very expensive to predict in terms of resources compared to depth wise convolutional model.

Table 2: Training and Validation accuracies

Model	**Training Accuracy**	**Testing Accuracy**
Sequential Model	84.22	64.62
Depth Wise Convolutional Model	67.75	60.95

3.3 Why Depth Wise Separable Convolutional Model Over Sequential Model? (Move table down)

Standard convolution is slow to perform operations. The computation required for convolutional can be measured by a number of multiplications required.

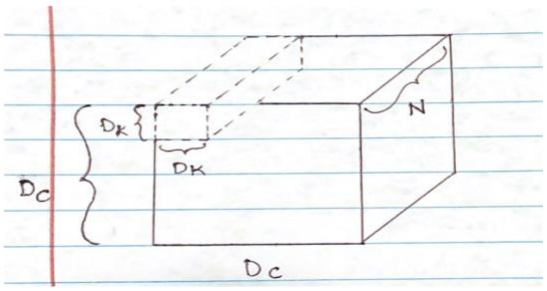

Figure 6. Input volume with Kernel

Consider the input of dimensions D_c X D_c X N, where N is the number of channels. If the input image is color, then N=3, and if it is grayscale, then N=2. D_K represents the kernel size. Convolutional operations take the sum of products of the input and the Kernel to return the scaler. For one convolutional operation, the number of multiplications required is D_c^2 X M. Since the Kernel is slide over input, the total number of convolutions over width and height is D_x^2. Hence a total number of multiplications of one Kernel over the input F is D_x^2 X D_K^2 X N. For L kernels; it will L X D_x^2 X D_K^2 X N.

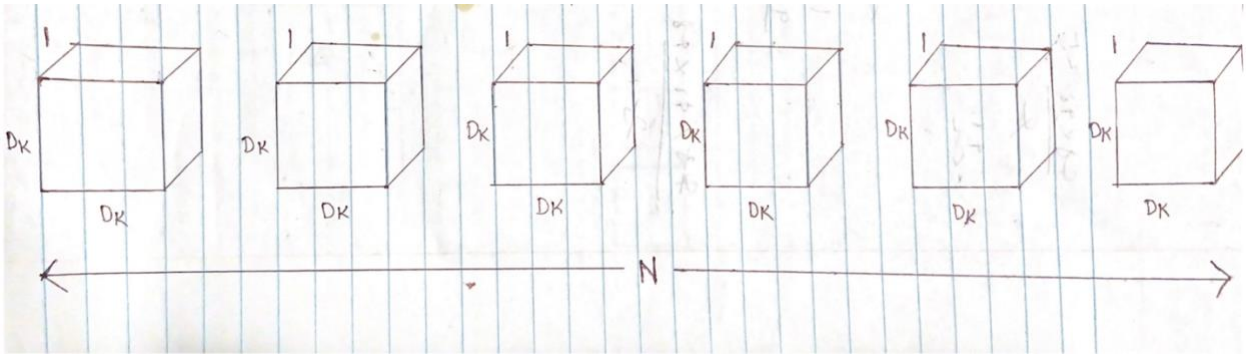

Figure 7. Kernel dimensions

Depth Wise Separable Convolutional applies convolution to a single channel at a time [18]. It is divided into two stages, namely depth-wise convolution and point-wise convolution. Consider Fig. 7; in Depth Wise Convolution, Fig.7 N Kernels of shape D_K X D_K X 1 are used. Here depth is one since it is only applied to one channel.

Since we apply one Kernel to a single input channel, we require M such kernels. After stacking all outputs for each of these N convolutions, we get an output volume of V with shape D_v X D_v X N.

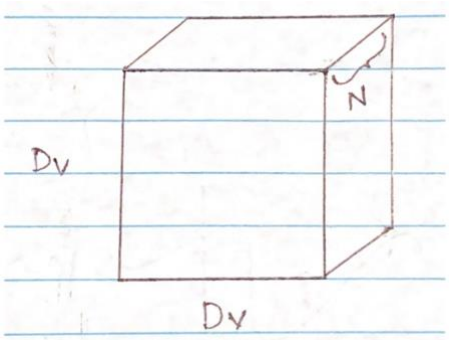

Figure 8. convolution output for M channels

In the point wise convolution, the linear convolution of each layer is performed [18]. Fig. 8 represents Input of volume D_v X D_v X N and filter has shape 1 X 1 X N. After performing 1X1 convolution operation over N layers and considering L filters, we get an output of shape of D_G X D_v X L. In the first Depth Wise Convolution stage, the number of multiplications required is N X D_v^2 X D_K^2 and in point wise convolution stage, the number of multiplications required is L X D_v X D_v X N. So, a total number of multiplications required are N X D_v^2 (D_K^2 + L). The ratio of multiplications of standard and depth wise convolution is ($1/L + 1/D_K^2$). If we consider L = 1024 and D_K = 3, the ratio obtained is 1/9, which means standard convolution requires 9 times more multiplications than depth wise convolution. That is one of the major reasons why we finalized depth wise convolution model for deployment.

As per Table 2, the sequential model has better accuracy than depth wise convolution model, but the sequential model is very heavy. The model file size is 43MB as compared depth wise convolution model, which is 975KB in size. After testing both the models on the Android application, the application had performance issues (application crash, delay in producing output, delay in importing the model) when using the sequential model. These issues were solved when we used Depth Wise Convolution Model.

3.4 Speech Emotion Recognition

As the fundamental research of affective computing, Speech Emotion Recognition (SER) has become an active research area [11]. Most of the SER frameworks comprise two sequential phases, one phase is feature extraction, and the next is emotion recognition. Human speech contains features like MFCC, chroma Short Time Fourier Transform, Mel spectrogram, and contrast using which useful information can be extracted to analyze the human behavior. tonnetz. H. M. Fayek, M. Lech, and L. Cavedon [12] performed few analyses over CNN and LSTM-RNN models. The CNN architectures achieved better performance and results as compared to other LSTMRNN designs. Z. HAN and J. WANG [13] have integrated features extracted from Deep Belief Networks (DBN) and traditional features as have used them to train the model. They have used six nonlinear Proximal Support Vector Machines (PSVM) to recognize the emotions. The majority out of six is considered as final emotion.

3.4.1 Model Architecture

The speech model contains one convolution 1D, four convolution 1D layers. The activation function used is relu.

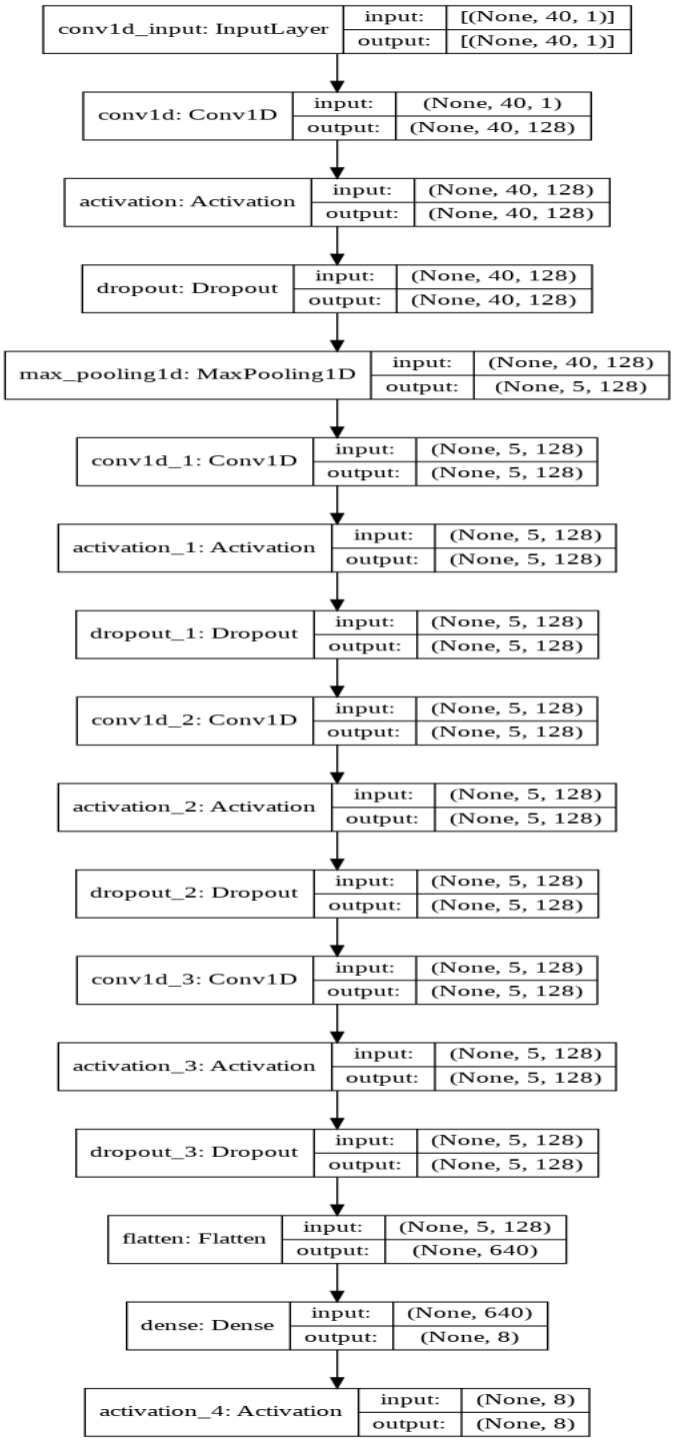

Figure 9. Speech Model Architecture

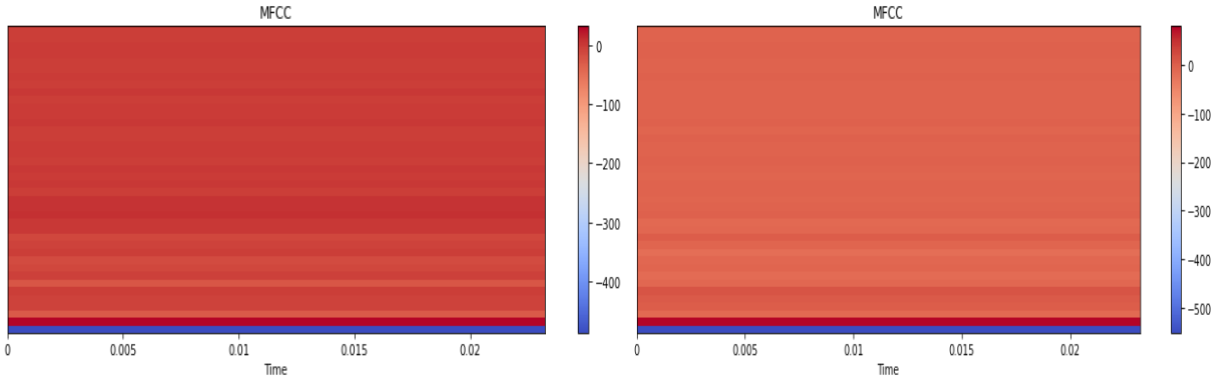

Figure 10. MFCC representation (Angry)

Figure 11. MFCC representation (Sad)

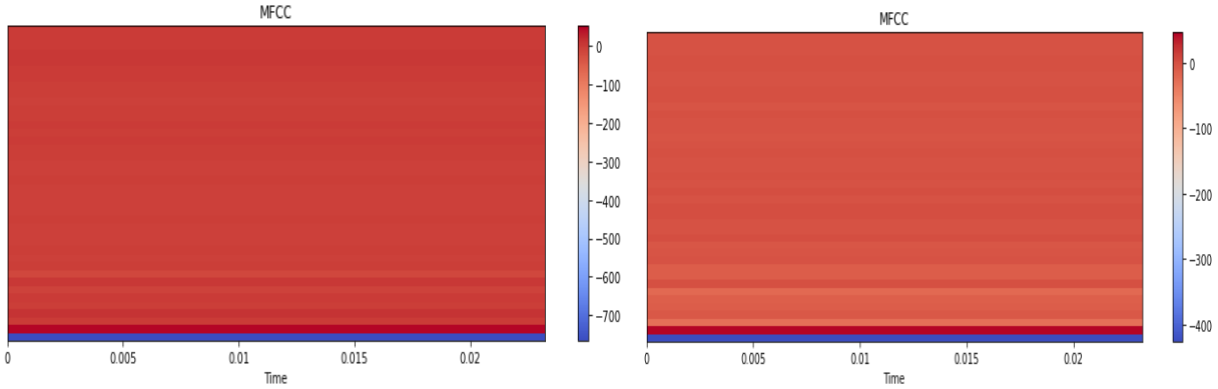

Figure 12. MFCC representation (Neutral)

Figure 13. MFCC representation (Happy)

As you can see from the MFCC representation, the MFCC of every emotion is represented by different color intensity. The angry coefficient is darker as compared to sad. Happy is lighter as compared to other emotions. Based on these MFCCs, our speech recognition model would be able to identify the emotions.

3.5 Multi-Sensory Emotion Recognition

A multi-sensory emotion recognition methodology is used to combine the output obtained from multiple sensory components (face, speech). We have implemented an Information Fusion algorithm [19] to perform this output integration.

3.6 Information Fusion

The emotion recognition process consists of information fusion, which refers to integrating and combining all the information from the user and representing it in terms of one emotion. i.e., integrating the facial expression emotion and speech emotion to calculate one single emotion. There are two ways by which we can do information fusion.

3.6.1 Feature level fusion

Feature level fusion is performed on the features extracted from the models (In our case speech, and facial expression features). These features are fused to find the dominant emotion. However, the fusion is computationally expensive and time-consuming in the case of high-dimensional data. Also, this feature level fusion is difficult to normalize since the features obtained are different in terms of characteristics.

3.6.2 Decision level fusion

Decision level fusion assumes that every module is independent of each other irrespective of their output (In our case, there are two modules, face, and speech). Each module is classified independently, and the result of both modules is unified to get the overall emotion of a conversation over a period. This level of fusion is also computationally inexpensive. Since we have limited computational resources in mobile, we have chosen the decision level information fusion in this project.

3.7 Fusion Algorithm [19]

This algorithm is used to associate the emotions obtained from the facial expression and speech model and predict the one emotion over the period.

1. Count the individual number of emotions given by the facial model over the period. For example, in 10 seconds, Happy came four times, neutral three times, and sad three times, etc.
2. Similarly, count the emotions given by the speech model over the period.
3. Store topmost emotion probability and the respective emotion gave by both facial and speech model. For example, according to (1) PF_{max} = 4/10 and F_{emo} = Happy. Similarly let us consider PS_{max} = 5/10 and S_{emo} = Neutral.
4. According to [2], during the communication between human beings, 55% of emotions and attitudes are represented by facial expression; ,the acoustic feature of speech shows 38% and 7% is related to the spoken words.
5. Assign a weight to F_{max} and S_{max} by multiplying W_f *PF_{max} and W_s *PS_{max}, where W_f = 0.55 and W_s = 0.38 are constant weights calculated based on (4).
6. Assign the Emo_{Final} as either F_{emo} or S_{emo} based on the max (W_f *$PF_{max,}$ W_s *PS_{max}).
7. Repeat steps 1 to 6.

The advantage of this algorithm is misclassification happened by one model (face or speech) can be covered by other. For example, consider PF_{max} = 3/10 and F_{emo} = Fear and PS_{max} = 5/10 and S_{emo} = Sad. Assuming overall emotion of conversation is Sad, if face emotion model is misclassifying the emotion and giving the wrong output, the final output will be Sad as max (0.55*(3/10), 0.38 *(5/10)) = 0.19 and Final Emotion is Sad. This indicates even if facial attributes has higher weightage, it won't affect the final output in case of misclassification.

IV. Android Application

The developed Android application can access two cameras simultaneously, and it is also able to access the mic to record the conversation. This application aims to understand the emotion behind communication between two or more people facing front and back cameras.

Figure 14. Android Application User Interface

The application is developed in Java. Originally, the facial and speech emotion recognition models were developed in Keras and trained and tested on Google Colab. Since devices like mobile have limited computing power, Keras models are converted into TensorFlow-lite models to increase the speed of the application. These lite models are deployed on the Android application to get the desired result.

4.1 Face Emotion Detection Process

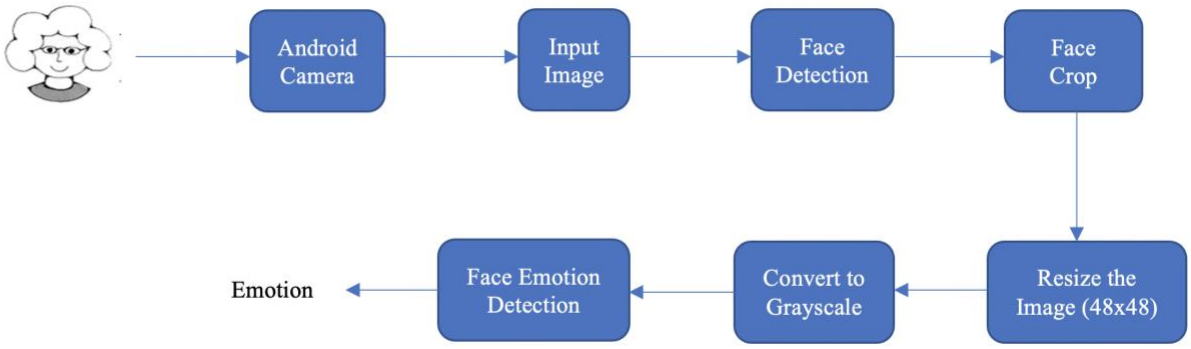

Figure 15. Emotion recognition process (Face)

For the facial expression part, both front and back cameras are capturing one image/frame sequentially. These images are passed to the Android facial detection library named Viola. Viola detects the face and returns the faces out of the images. The newly obtained face images are normalized and converted into grayscale. Those grayscale images are converted into Tensor-Image and reshaped into 48 X 48 using Bilinear Image resize technique and passed to the model. The model outputs an array containing the probability of 7 emotions to which that image is classified. The emotion with the highest probability is the emotion of that frame.

4.2 Speech Emotion Detection Process

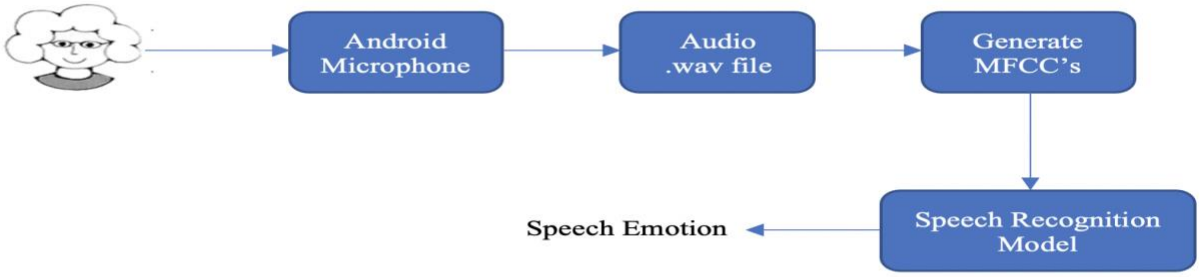

Figure 16. Emotion recognition process (Speech)

Similarly, for the speech recognition part, the application records the conversation between two people facing the front and back camera every 10 seconds. The recording is stored in WAV [17] format. Since Android does not provide inbuilt support to record files in WAV [17] format, we use an external Android dependency named RecordWav. The recorded WAV [17] file is used to generate Mel Frequency Cepstral Coefficient (MFCC). Android provides JLibrosa library support to extract MFCCs from WAV [17]. The calculated MFCC are then passed to the speech emotion model to predict the emotion of the conversation. The process is repeated after every 10 seconds.

The application is built in Android studio and has been tested on google pixel. The application is hardware-dependent since companies like Samsung do not allow access to both cameras simultaneously. The application has been developed by implementing various Android core programming concepts like multi-threading and async tasks.

V. Dataset

5.1 Face Expressions Dataset

For training the facial emotion recognition model, we have used the FER-2013 dataset. The dataset contains 35,587 Images of people displaying various emotions through facial expressions.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 35887 entries, 0 to 35886
Data columns (total 3 columns):
 #   Column   Non-Null Count  Dtype 
---  ------   --------------  ----- 
 0   emotion  35887 non-null  int64 
 1   pixels   35887 non-null  object
 2   Usage    35887 non-null  object
dtypes: int64(1), object(2)
memory usage: 841.2+ KB
```

28,709 random images were used for model training, and 7178 images were used for testing purposes. The images can be classified into 7 types of emotions: Happy, Sad, Angry, Fear, Disgust, Surprise, and Neutral. For the preprocessing, images were converted into grayscale and then reshaped into 48x48 dimensions.

Figure 17. Images after Preprocessing

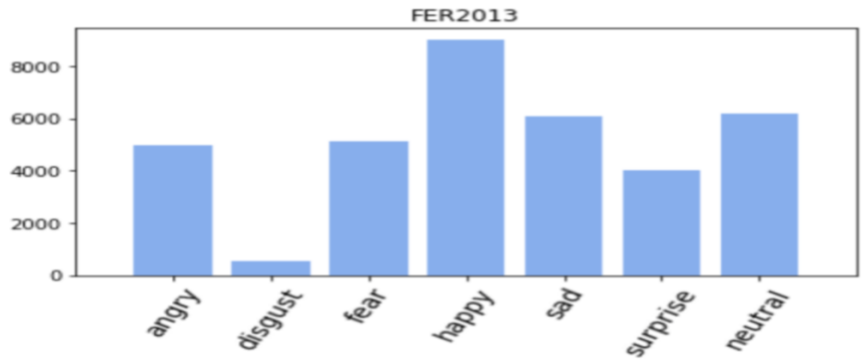

Figure 18. Distribution of the number of images per emotion

5.2 Speech Emotion Dataset

For emotion recognition based on speech, we have used the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS). The dataset contains a total of 605 audio recordings expressing different emotions using a speech by 24 different actors. We have divided data into 75% and 25% as training and testing, respectively. The audio format is WAV [17]. The recordings can be classified into four classes, namely Happy, Sad, Neutral, and Angry.

The 40 MFCCs are extracted from the audio files. These extracted MFCCs are used to train the speech model and classify. Further model is tested on multiple features like chroma and Mel spectrogram. If we use all these features, the total number of coefficients becomes 182.

VI. Experiments

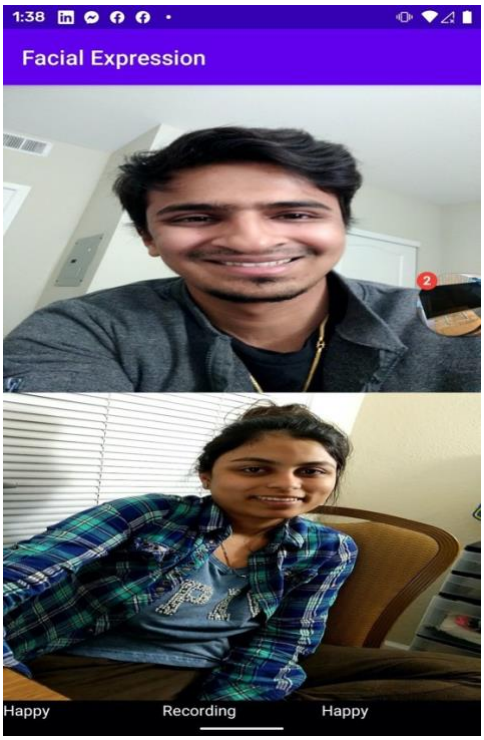

Figure 19. Facial Emotion Recognition

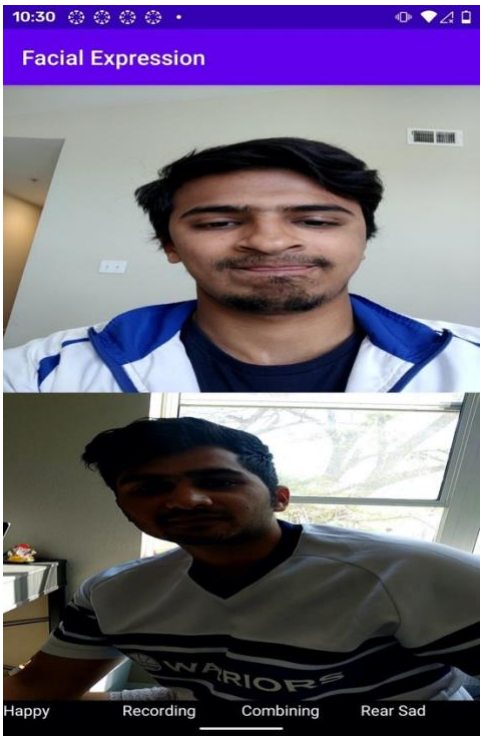

Figure 20. Emotion with facial and speech features

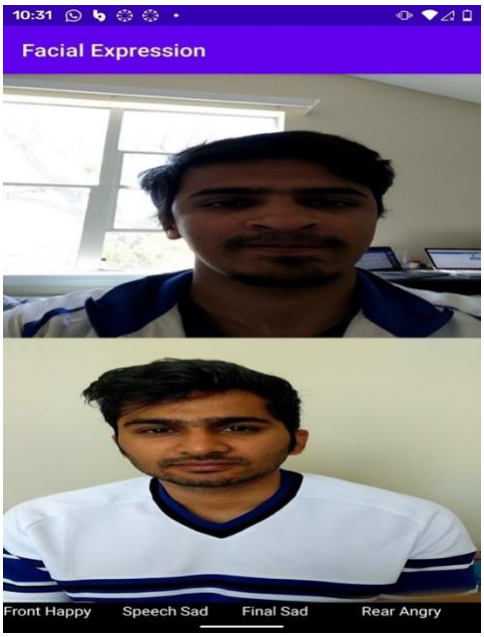

Figure 21. Combined result with facial and speech features

Figure 22. Image after face detection and preprocessing

Fig. 19 shows the initial phase of the model where it can detect emotions based on facial expressions. As you can see, both persons facing the camera are smiling; hence the model has predicted the emotions as Happy. Fig. 20 and Fig. 21 are classic examples of misclassification. The front camera person is not smiling or does not seem happy, but the model has predicted its emotion as happy. But the rear camera emotions are up to the mark. In Fig. 20, a person facing the rear camera looks sad, and the application is still recording the conversation between two people. In Fig. 21, the person facing the rear camera looks somewhat neutral and angry, and the model has predicted it correctly. The speech emotion predicted is sad, and the final emotion calculated using the Fusion Algorithm is also sad. In Fig. 22, we can observe the output image which we are providing to the model. The face is detected from the image, and it is reshaped and converted into grayscale.

6.1 Results: Facial Emotion Recognition

6.1.1 Sequential Model

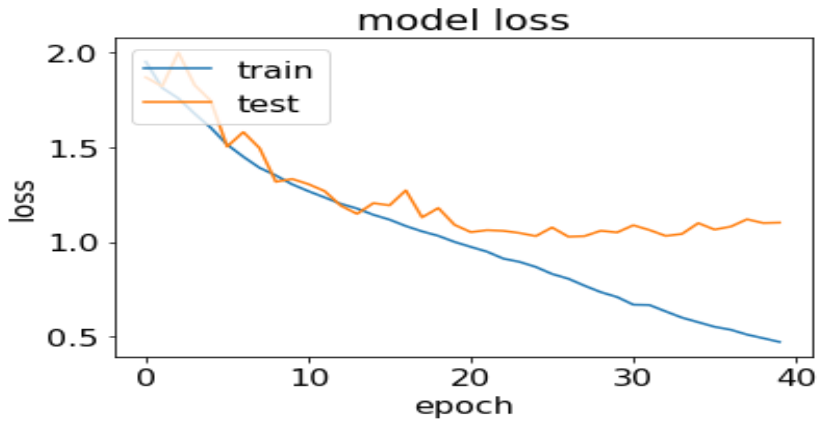

Figure 23. Training and Validation loss (Sequential Model)

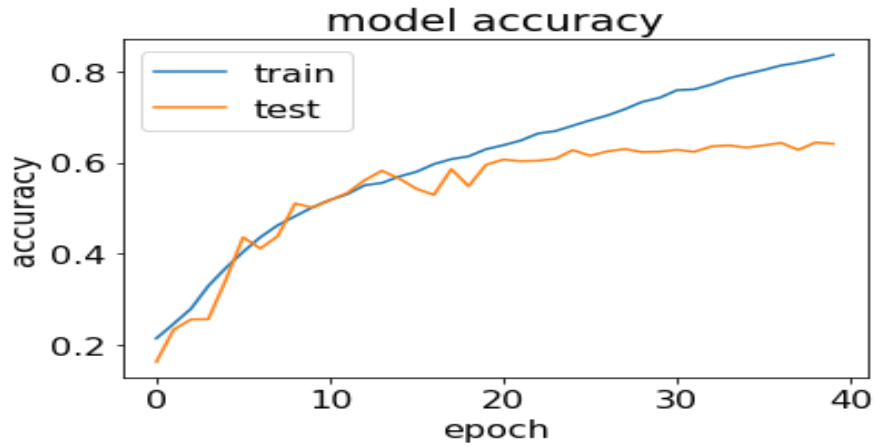

Figure 24. Training and Validation accuracy (Sequential Model)

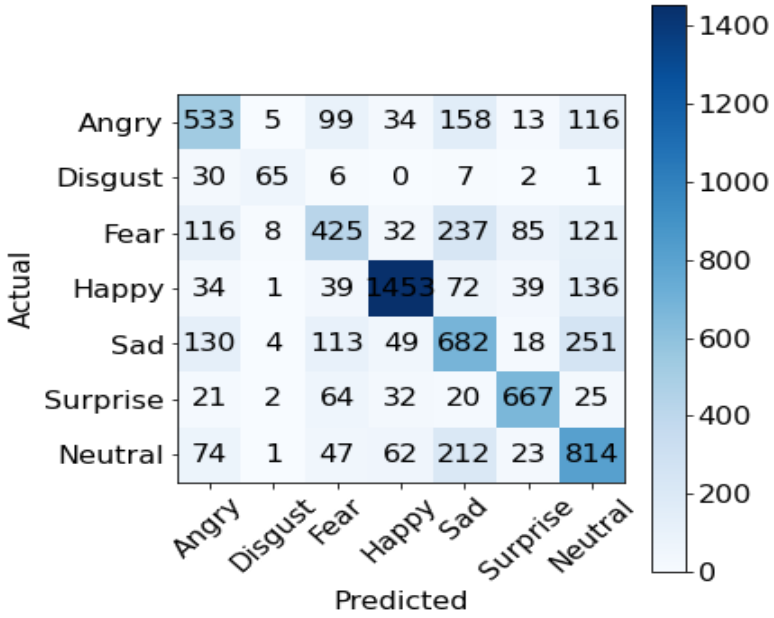

Figure 25. Confusion Matrix (Sequential Model)

As you can see in the training and validation loss, training loss is decreasing, but the validation loss has started to increase after the 37th-38th epoch. Happy is the best-classified emotion with an accuracy of 81.90%, and fear is the worst classified emotion with an accuracy of 41.50%. Fear emotion is often misclassified into Sad, Surprise, and Neutral. We assume this has happened

since all these emotions have minor changes while expressing it on the face. You can also see that Sad is majorly misclassified into Neutral because of the same reason. Overall testing accuracy of the model 64.62%.

6.1.2 Depth Wise Convolution Model

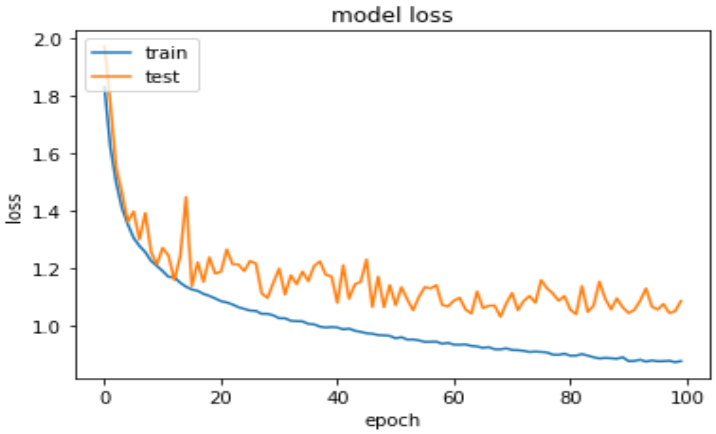

Figure 26. Training and Validation loss (Depth wise Model)

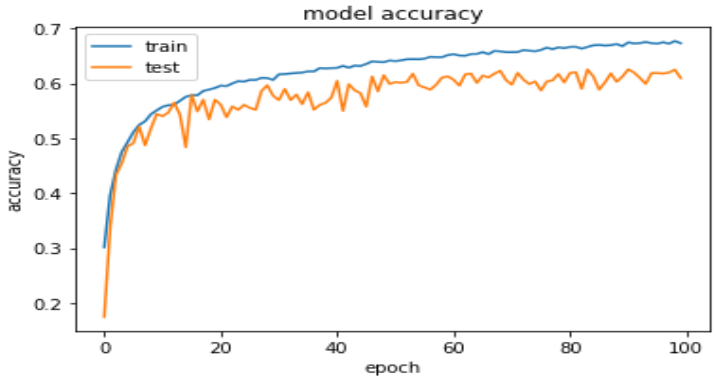

Figure 27. Training and Validation accuracy (Depth wise Model)

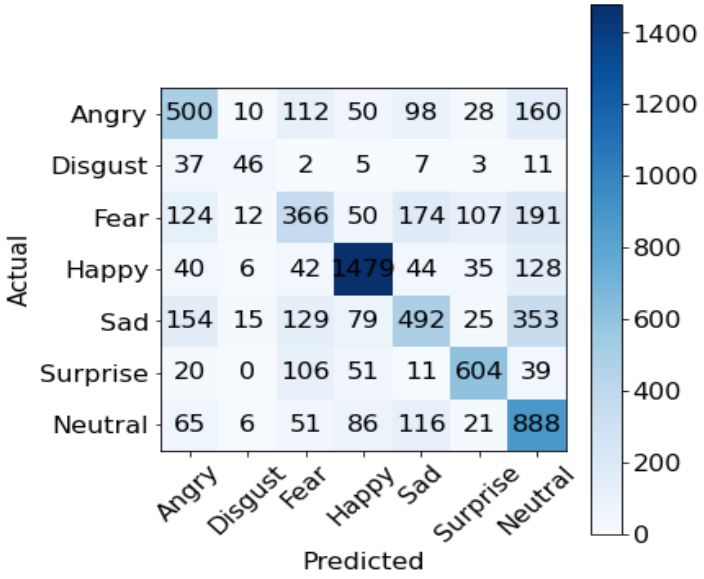

Figure 28. Confusion Matrix (Depth wise Model)

Here we can see that model is a good fit as training and validation loss curves are going down, and the time required to train the model is also less as compared to the sequential model. The accuracy curve is also similar and going up. Here as well, Happy is the best-classified emotion with an accuracy of 83.37%, and fear is the worst classified emotion with an accuracy of 35.74%. As you can see, fear is majorly misclassified into Angry, Sad, and Neutral. Overall testing accuracy is 60.95%.

6.2 Results: Speech Emotion Recognition

We tested the Speech Recognition model using multiple features like MFCCs, Chroma STFT, and Mel Spectrogram. As we increased the features, we observed a minor increase in the accuracy. But at the same time, an increase in the size of the input was observed.

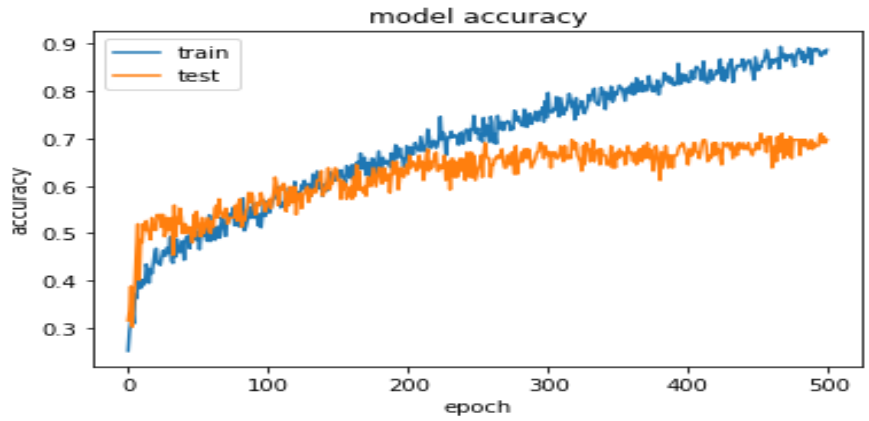

Figure 29. Training and Validation Accuracy all features (Speech)

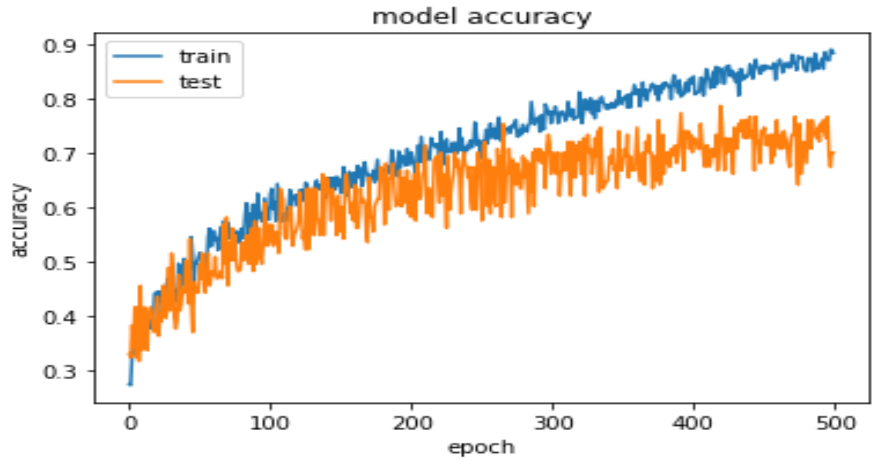

Figure 30. Training and Validation Accuracy MFCC and Mel Spectrogram (Speech)

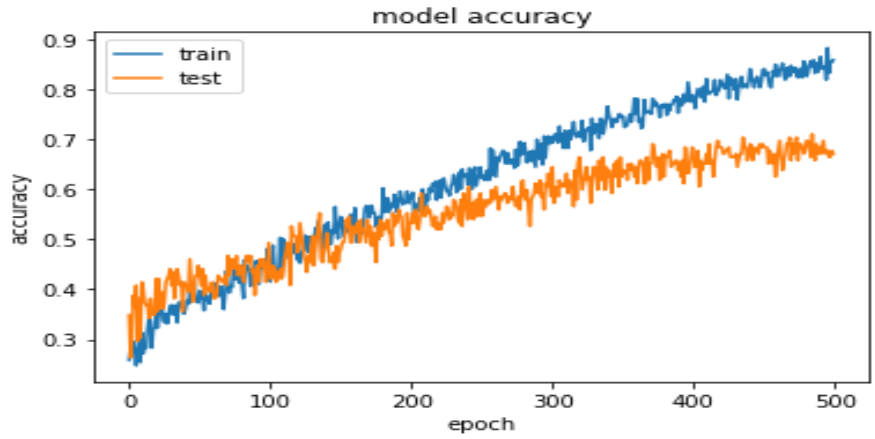

Figure 31. Training and Validation Accuracy MFCC (Speech)

As you can see from the accuracy graphs, as we are decreasing the number of features considered, the accuracy is decreasing. But it is not a significant decrease. Hence ultimately, it comes to the accuracy vs size of the input. We have only used MFCC for model prediction.

6.2.1 Why have we used only MFCCs as a feature for Speech Emotion Prediction?

In python, there is a library named Librosa, which is used to calculate all these features (MFCCs, Chroma, and Mel Spectrogram). For Android, there is a substitute library named JLibrosa, which provides similar features. As this library is still under development, the library does not provide the functionality to calculate Chroma and Mel Spectrograms. If we consider all three features, the input size is (182,1), which is large as compared to considering only MFCC where input size (40,1). The difference between accuracies is not significant, and as per the application requirement, the time taken to predict the emotion should be less. Hence, we are considering only MFCC to predict the emotion.

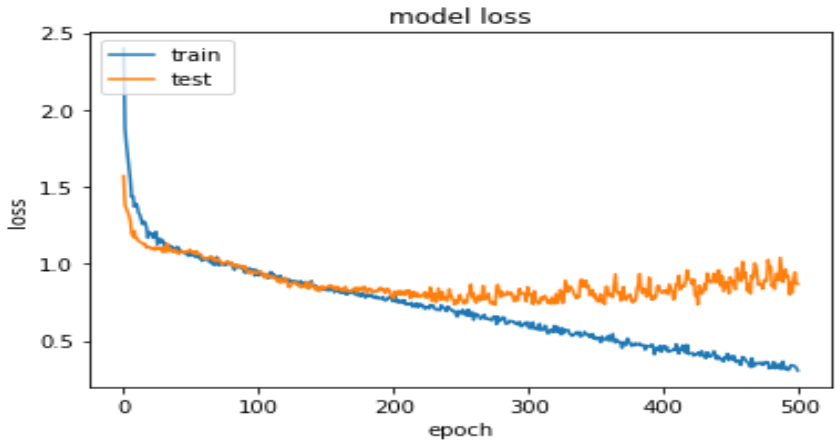

Figure 32. Training and Validation loss all features (Speech)

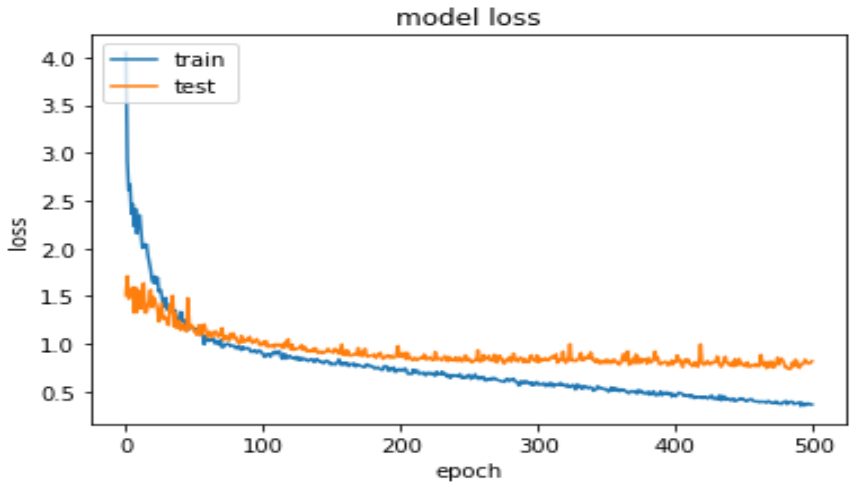

Figure 33. Training and Validation loss MFCC and Mel Spectrogram (Speech)

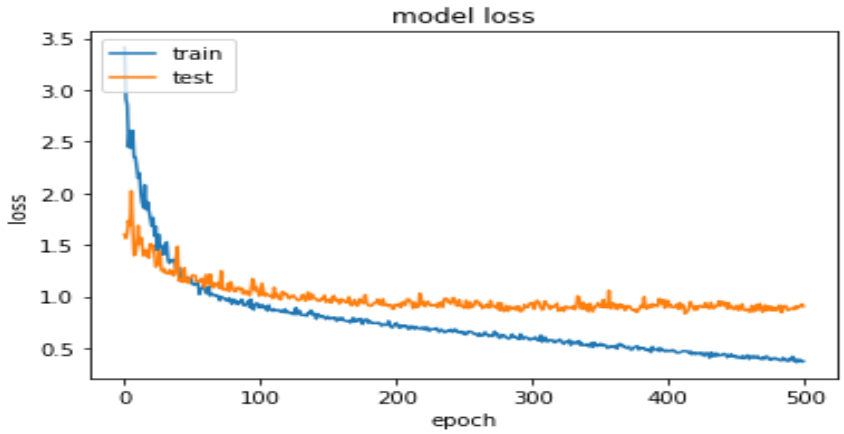

Figure 34. Training and Validation loss MFCC (Speech)

As we can see in the train and test loss graph, the loss is decreasing as the epochs are increasing. There is no significant difference in the loss when we use only MFCC to train the model instead of all three features (MFCC, chroma, and Mel Spectrogram).

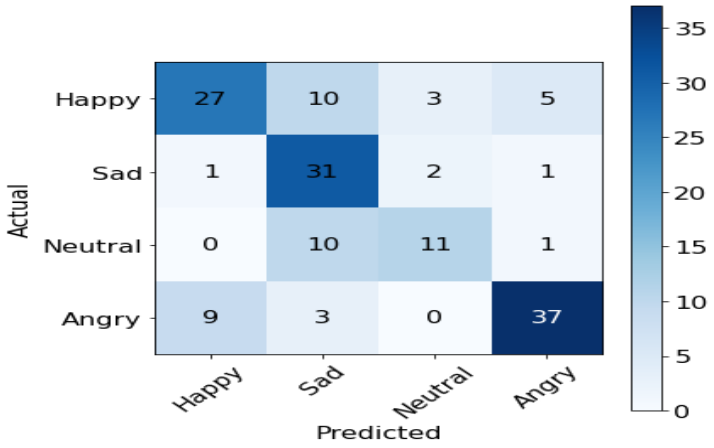

Figure 35. Confusion Matrix (Speech)

For the Speech Emotion Recognition part, we have used four classes to identify emotions like Fear, Surprise, and Disgust are difficult to produce and understand by human speech. As we can observe from the confusion matrix, Sad is the best-classified emotion with an accuracy of 88%, and Neutral is the worst classified emotion with an accuracy of 50%. The reason behind it is the pitch of the speech. By estimating this result, we anticipated that every human being has a different sound pitch while expressing emotions. Hence someone's Neutral pitch can be another person's Sad pitch. Hence there is a high misclassification in Neutral Emotion.

VII. Conclusion and Future Work

Emotion recognition using deep learning is one of the most researched problems. There are many techniques by which human emotion can be identified. But those techniques do not give promising results if used individually. Multimodal is one of the ways that looks promising. As in our project, we have used two techniques, namely speech emotion, and facial emotion recognition, simultaneously. To find one final emotion, we have used a fusion algorithm to fuse the output received from both models. One thing observed is classic machine learning techniques do not do well on images and audio datasets in terms of accuracy. Hence, we have used sequential and depth-wise-convolutional models for facial expression and speech, respectively. The tflite models are deployed on an Android application that can access both cameras simultaneously and can record the conversation. The weights (W_f and W_s) are added to the output based on Albert communication model (Mehrabian, 1971). The facial emotion accuracy and the weightage in emotion representation are more as compared to speech emotion.

Based on our research and development, additional opportunities for future research in this field are possible to amplify the current 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. Another improvement to this can be to implement more human attributes into the application, such as hand gesture and body gesture and speech text, but this can be challenging as we have limited computing resources in mobile. The UI and the performance of the Android application can be improved as well. Since JLibrosa is at its development stage, it only gives us the functionality to calculate MFCCs. Further, after enhancement, we can include more speech features like chroma, Mel spectrogram, contrast to train the speech model and increase its accuracy. This application with multimodal concept can be useful for the people whose face is unfortunately damaged. In that case,

speech model will cover the inaccuracy of face emotion recognition. Also, people who are facing Autism problem develops a disorder that impairs the ability to communicate or interact. This application can also be useful to understand their emotions.

References

[1] X. Zhang, M. -J. Wang and X. -D. Guo, "Multimodal Emotion Recognition Based on Deep Learning in Speech, Video and Text," 2020 IEEE 5th International Conference on Signal and Image Processing (ICSIP), Nanjing, China, 2020, pp. 328-333, doi: 10.1109/ICSIP49896.2020.9339464.

[2] 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.

[3] S. K. Pandey, H. S. Shekhawat and S. R. M. Prasanna, "Deep Learning Techniques for Speech Emotion Recognition: A Review," *2019 29th International Conference Radioelektronika (RADIOELEKTRONIKA)*, Pardubice, Czech Republic, 2019, pp. 1-6, doi: 10.1109/RADIOELEK.2019.8733432.

[4] L. B. Letaifa, M. I. Torres and R. Justo, "Adding dimensional features for emotion recognition on speech," 2020 5th International Conference on Advanced Technologies for Signal and Image Processing (ATSIP), Sousse, Tunisia, 2020, pp. 1-6, doi: 10.1109/ATSIP49331.2020.9231766.

[5] The 6 Types of Basic Emotions and Their Effect on Human Behavior verywellmind.com/an-overview-of-the-types-of-emotions-4163976

[6] E. Ghaleb, M. Popa and S. Asteriadis, "Metric Learning-Based Multimodal Audio-Visual Emotion Recognition," in *IEEE MultiMedia*, vol. 27, no. 1, pp. 37-48, 1 Jan.-March 2020, doi: 10.1109/MMUL.2019.2960219.

[7] D. Kalita, "Designing of Facial Emotion Recognition System Based on Machine Learning," *2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*, 2020, pp. 969-972, doi: 10.1109/ICRITO48877.2020.9197771.

[8] Du, Shichuan, and Aleix M Martinez. “The resolution of facial expressions of emotion.” *Journal of vision* vol. 11,13 24. 30 Nov. 2011, doi:10.1167/11.13.24

[9] G. Luh, H. Wu, Y. Yong, Y. Lai and Y. Chen, "Facial Expression Based Emotion Recognition Employing YOLOv3 Deep Neural Networks," *2019 International Conference on Machine Learning and Cybernetics (ICMLC)*, 2019, pp. 1-7, doi: 10.1109/ICMLC48188.2019.8949236.

[10] Qingyang Zhu, Guanming Lu, and Jingjie Yan. 2020. Valence-Arousal Model based Emotion Recognition using EEG, peripheral physiological signals and Facial Expression. In Proceedings of the 4th International Conference on Machine Learning and Soft Computing (ICMLSC 2020). Association for Computing Machinery, New York, NY, USA, 81–85. DOI:https://doi-org.libaccess.sjlibrary.org/10.1145/3380688.3380694

[11] S. Ramakrishnan and I. Emary, "Speech emotion recognition approaches in human computer interaction," Telecommunication Systems, vol. 52, pp. 1467–1478, 2013.

[12] H. M. Fayek, M. Lech, and L. Cavedon, "Evaluating deep learning architectures for Speech Emotion Recognition," Neural Networks, 2017, doi:10.1016/j.neunet.2017.02.013.

[13] Z. HAN and J. WANG, "Speech Emotion Recognition Based on Deep Learning and Kernel Nonlinear PSVM," 2019 Chinese Control And Decision Conference (CCDC), 2019, pp. 1426-1430, doi: 10.1109/CCDC.2019.8832414.

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

[15] Leland Roberts, Understanding the Mel Spectrogram https://medium.com/analytics-vidhya/understanding-the-mel-spectrogram-fca2afa2ce53

[16] Ferris Jabr, The evolution of emotion: Charles Darwin's little-known psychology experiment https://blogs.scientificamerican.com/observations/the-evolution-of-emotion-charles-darwins-little-known-psychology-experiment/#:~:text=In%201872%2C%20Darwin%20published%20The,emotion%20through%20remarkably%20similar%20behaviors.&text=By%20stimulating%20the%20right%20combination,Duchenne%20mimicked%20genuine%20emotional%20expression.

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

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

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

[20] Pratheeksha Nair, The dummy's guide to MFCC https://medium.com/prathena/the-dummys-guide-to-mfcc-aceab2450fd

[21] Cindi May, “Are Women More Emotionally Expressive Than Men?” https://www.scientificamerican.com/article/are-women-more-emotionally-expressive-than-men/