International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Emotion Recognition based on Speech and Facial Expression

Mrunmayi Ramji Patankar[1]*, Harshita Rajesh Madane*[2]*, Insha Afaque Mulla*[3]

[1,2,3]Department of Information Technology Vidyavardhini's College of Engineering and Technology Vasai , India

ABSTRACT—

Human feeling is the best to understand the atti- tude of anyone towards another or a particular situation. The fundamental motive is to teach a machine about the emotions that humans possess, which is also an essential demand within the domain of social intelligence, simultaneously speeding up the progress of human-to-machine and machine-to-human interac- tions.In general, a person's emotions are a combination of speech and facial expressions. The face may be an advanced interaction of assorted activities, displaying human emotions expressed by many signals. The facial expressions will describe one neutral and 5 basic emotions - happiness, sadness, fear, surprise, and anger. The computers of future generations should be ready to move with a person being similar to another.Our model intentionally targets the user base who are failing to specify it or are finding it problematic. Our aim is to develop a model that's ready to offer a live sentiment analysis with a visible program. Therefore, we have set the video type of inputs: Video input from a live digital camera or keep from an associate degree MP4 file, from which we tend to split the audio and therefore the pictures. Will use face markers. In conjunction with simultaneous speech recordings, detailed facial motions are captured with motion capture.

International Journal of Research Publication and Reviews, Vol 4, no 4, pp 3505-3509, April 2023

Index Terms—**Bimodal, Emotions, Speech, Image, Recognition, OpenCV, machine learning, image processing, neural network.**

I. INTRODUCTION

In many different industries around the globe, emotion recognition is extremely important. Automatic facial feeling recognition has drawn a lot of interest over the past 20 years. This is frequently due to the increase in demand for behavioural biometric systems and human-machine interac- tion, in which facial emotion detection and, consequently, the intensity of emotion, play key roles.

We have demonstrated a model developed for facial feature recognition, using various libraries. Similarly, speech may be an advanced signal, therefore from the speech, we classified a proof feeling by taking some options like pitch, entropy, and energy. These speech options are language-independent and non-verbal. The audio extractor mechanically filters out the audio signal from the captured video and audio signals of one person. Currently, the video is split into frames. The emotion of each frame is recognized. The emotion that seems in most of the frames currently becomes the video signal's emotion. Similarly, the audio signal's emotion is classed.Our project aims to recognize human emotions with the help of speech as well as image. The solution includes various deep learning algorithms, image processing, and neural networks. For speech, we have used CNN and LSTM to build our model. We have tried various combinations to check the accuracy. The frame rate for speech recognition is 20 fps and 40 is specified as a coefficient. The frame size for speech is 1 second.

II. TERMINOLOGIES

Here, are the few terminologies that first need to be under- stand before deep dive into the emotion recognition model.

A. *Deep Learning and its algorithms*

Artificial neural networks are used in deep learning to carry out complex calculations on vast quantities of data. It is a form of artificial intelligence that is founded on how the human brain is organised and functions. Machines are trained using deep learning algorithms by learning from instances. Deep learning is frequently used in sectors like healthcare, e-commerce, amusement, and advertising.

B. *Neural Networks*

Deep learning methods are based on neural networks, also referred to as artificial neural networks (ANNs) or simulated neural networks (SNNs), which are a subset of machine learning. Their structure and name are modelled after the human brain, imitating the communication between biological neurons.

C. *Feature Extraction*

Feature extraction is a process of dimensionality reduction by which an initial set of raw data is reduced to more manageable groups for processing. A characteristic of these large data sets is a large number of variables that require a lot of computing resources to process. Feature extraction is the name for methods that select and /or combine variables into features, effectively reducing the amount of data that must be processed, while still accurately and completely describing the original data set.

D. Model

Our model detects emotions on a real-time basis. Once the code is in a running state, it opens the webcam of the system and whatever expression is being displayed (in terms of facial expression and speech) is detected. It detects the emotions of a number of people at the same time as well. For an instance - if there are 3 people looking at the camera and speaking then it will simultaneously detect the emotions of all three of them. It keeps on detecting in milliseconds. A machine learning model is a file that has been trained to recognize certain types of patterns. You train a model over a set of data, providing it an algorithm that it can use to reason over and learn from those data.

III. FEATURE EXTRACTION

The following techniques are used to extract features from the speech:

A. Mel-Frequency Cepstral Coefficients (MFCC)

The Mel-Frequency Cepstral Coefficients (MFCC) tech- nique aims to develop the features from the audio signal which can be used for detecting the phones in the speech.It is a 1D technique, hence we do not have to make conversions. The formula used to calculate the mels for any frequency is:

$mel(f) = 2595xlog10(1 + f/700)$ (1)

where mel(f) is the frequency (mels) and f is the frequency (Hz). The MFCCs are calculated using this equation:

$Cn = kn = 1(logSk)cos[n(k12)k]$ (2)

where k is the number of mel spectrum coefficients, S k is the output of filterbank and C n is the final mfcc coefficients. The block diagram of the MFCC processor can be seen in the figure number **??** and it appears on page number 2. We have obtained 20 mfcc coefficients for each 1 second audio and here is the sample coefficients that we have obtained for 1 audio can be seen in the figure number 2

B. Linear Prediction Coefficients (LPC)

Linear Prediction Coefficients (LPC) imitate the human vocal tract and gives robust speech feature. It evaluates the speech signal by approximating the formants, getting rid of its effects from the speech signal, and estimating the concen- tration and frequency of the left-behind residue. The result states each sample of the signal is a direct incorporation of previous samples. The coefficients of the difference equation characterize the formants, thus, LPC needs to approximate these coefficients. LPC is a powerful speech analysis method and it has gained fame as a formant estimation method. Linear prediction analysis of speech signal forecasts any given speech sample at a specific period as a linear weighted aggregation of preceding samples. The linear predictive model of speech creation is given as:

$s(n) = pk = 1aks(nk)$ (3)

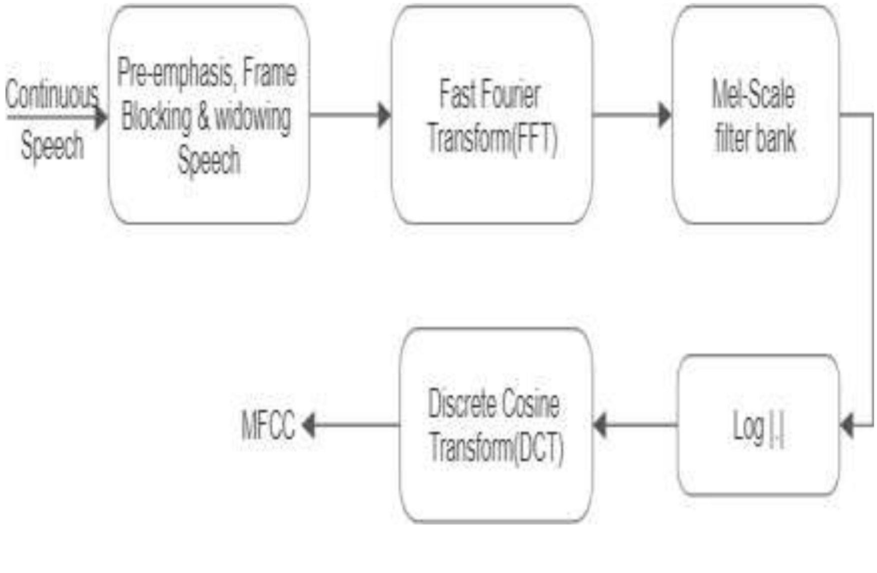
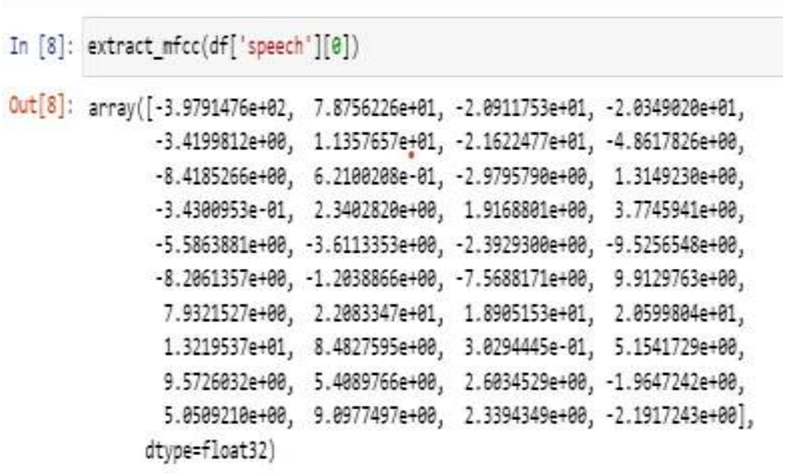

Continuous Speech
Pre-emphasis, Frame Blocking & widowing Speech
Fast Fourier Transform(FFT)
Mel-Scale filter bank
Log |.|
Discrete Cosine Transform(DCT)
MFCC

Fig. 1. Block diagram of MFCC processor.



```
In [8]: extract_mfcc(df['speech'][0])

Out[8]: array([-3.9791476e+02,  7.8756226e+01, -2.0911753e+01, -2.0349020e+01,
               -3.4199812e+00,  1.1357657e+01, -2.1622477e+01, -4.8617826e+00,
               -8.4185266e+00,  6.2100208e-01, -2.9795790e+00,  1.3149230e+00,
               -3.4300953e-01,  2.3402820e+00,  1.9168801e+00,  3.7745941e+00,
               -5.5863881e+00, -3.6113353e+00, -2.3929300e+00, -9.5256548e+00,
               -8.2061357e+00, -1.2038866e+00, -7.5688171e+00,  9.9129763e+00,
                7.9321527e+00,  2.2083347e+01,  1.8905153e+01,  2.0599804e+01,
                1.3219537e+01,  8.4827595e+00,  3.0294445e-01,  5.1541729e+00,
                9.5726032e+00,  5.4089766e+00,  2.6034529e+00, -1.9647242e+00,
                5.0509210e+00,  9.0977497e+00,  2.3394349e+00, -2.1917243e+00],
              dtype=float32)
```

Fig. 2. MFCC Coefficients of single audio

where s is the predicted sample, s is the speech sample, p is the predictor coefficients. The prediction error is given as:

$e(n) = s(n)s(n)$ (4)

Subsequently, each frame of the windowed signal is au- tocorrelated, while the highest autocorrelation value is the order of the linear prediction analysis. This is followed by the LPC analysis, where each frame of the autocorrelations is converted into LPC parameters set which consists of the LPC coefficients. LPC can be derived by:

$am = log[1km/1 + km]$(5)

The block diagram of the LPC processor can be seen in the figure number 3 and it appears on page number 3. It is a 2D technique and speech is 1D, hence we have to make conversions. We have obtained 40 LPC coefficients for each 1 second audio and here is the sample coefficients that we have obtained for 1 audio can be seen in the figure number 4

C. Mel-spectrogram

The Mel-spectrogram is an effective technique to extract hidden features from audio and visualize them as an image. Unlike LPC, Mel-spectrogram is a 1D technique, hence con- versions are not required. We have obtained more than 20 Mel- spectrogram coefficients for each 1 second audio and here is the sample coefficients that we have obtained for 1 audio can be seen in the figure number 5

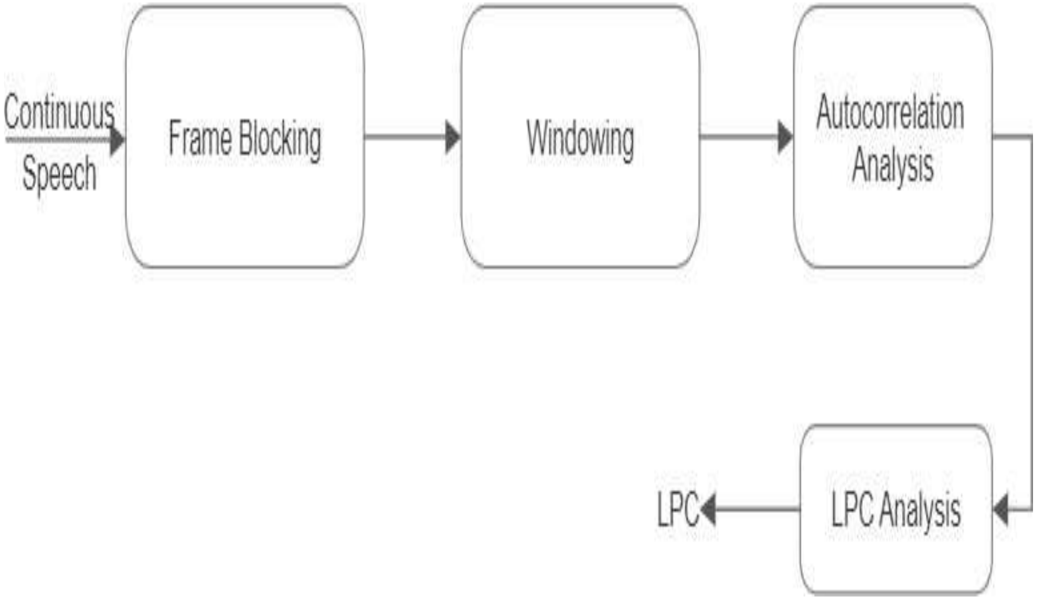
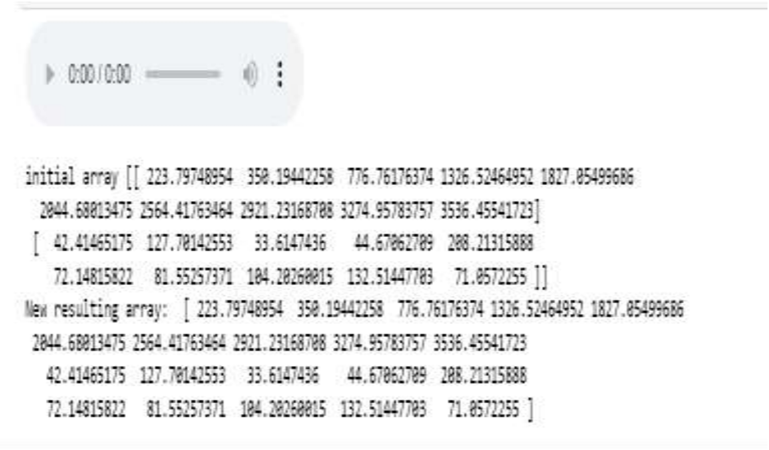
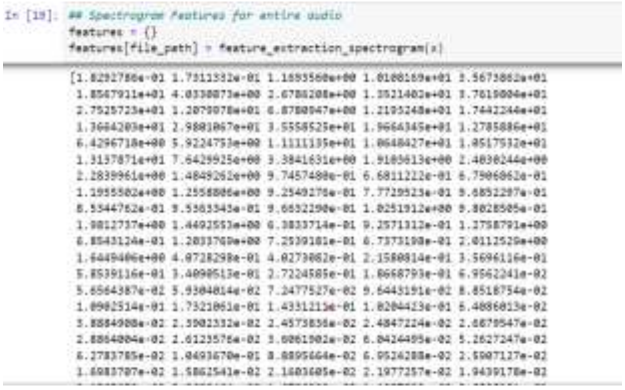


Fig. 3. Block diagram of LPC processor.



```
initial array [[ 223.79748954  350.19442258  776.76176374 1326.52464952 1827.05499686
  2044.68013475 2564.41763464 2921.23168708 3274.95783757 3536.45541723]
 [  42.41465175  127.70142553   33.6147436    44.67062709  208.21315888
    72.14815822   81.55257371  104.20260015  132.51447703   71.0572255 ]]
New resulting array:  [ 223.79748954  350.19442258  776.76176374 1326.52464952 1827.05499686
 2044.68013475 2564.41763464 2921.23168708 3274.95783757 3536.45541723
   42.41465175  127.70142553   33.6147436    44.67062709  208.21315888
   72.14815822   81.55257371  104.20260015  132.51447703   71.0572255 ]
```

Fig. 4. LPC Coefficients of single audio



```
In [19]: ## Spectrogram Features for entire audio
         features = {}
         features[file_path] = feature_extraction_spectrogram(x)

         [1.8292786e-01 1.7311332e-01 1.1693560e+00 1.0108169e+01 3.5673862e+01
          1.8567911e+01 4.0330873e+00 2.6788208e+00 1.3521402e+01 3.7619004e+01
          2.7525723e+01 1.2079078e+01 6.8780947e+00 1.2195248e+01 1.7442244e+01
          1.3664203e+01 2.9881067e+01 3.5558525e+01 1.9664345e+01 1.2785886e+01
          6.4296718e+00 5.9224753e+00 1.1111135e+01 1.0648427e+01 1.8517532e+01
          1.3137871e+01 7.6429925e+00 3.3841631e+00 1.9103613e+00 2.4030244e+00
          2.2839961e+00 1.4849262e+00 9.7457480e-01 6.6811222e-01 6.7906862e-01
          1.1955502e+00 1.2558806e+00 9.2549276e-01 7.7729523e-01 9.6852287e-01
          8.5344762e-01 9.5363343e-01 9.6652290e-01 1.0251912e+00 9.8028505e-01
          1.9812737e+00 1.4492553e+00 6.3833714e-01 9.2571312e-01 1.2758791e+00
          6.8543124e-01 1.2033769e+00 7.2539181e-01 6.7373198e-01 2.0112529e+00
          1.6449406e+00 4.0728298e-01 4.0273062e-01 2.1580814e-01 3.5696116e-01
          5.8539116e-01 3.4090513e-01 2.7224585e-01 1.0668793e-01 6.9562241e-02
          5.6564387e-02 5.9304014e-02 7.2477527e-02 9.6443191e-02 8.8518754e-02
          1.0902514e-01 1.7321061e-01 1.4331213e-01 1.0204423e-01 6.4886013e-02
          3.8884908e-02 2.3902332e-02 2.4573836e-02 2.4847224e-02 2.6679547e-02
          2.8864004e-02 2.6123576e-02 3.6061902e-02 6.0424495e-02 5.2627247e-02
          6.2783785e-02 1.0493670e-01 8.8895664e-02 6.9524288e-02 2.5907127e-02
          1.6983707e-02 1.5862541e-02 2.1603605e-02 2.1977257e-02 1.9439178e-02
```

Fig. 5. Mel-Spectrogram Coefficients of single audio

IV. MODELS

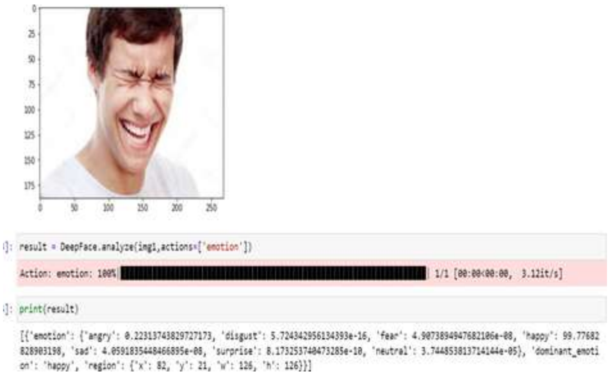
We've implemented 2 models: one for Facial Expression Recognition FACIAL EXPRESSION RECOGNITION and SPEECH EMOTION RECOGNITION to check the accuracy of our project.

A. *Facial Expression Recognition*

Through facial expressions, human emotion can be best shown. Classification of facial expressions from images can be built axiomatically from Facial expression recognition meth- ods. This system will be able to classify one neutral emotion, and five fundamental emotions including Happiness, Anger, Fear, Sadness, and Surprise. The facial expression recognition model has three stages. Here we have downloaded the dataset from kaggle(fer2013). The frames kept is 25fps.After training the model, accuracy achieved for CNN model was 66 percent.

– Face detection: For implementing face detection we have made use of the haracassade library and OpenCV2. – Feature extraction: For Extracting the features we have made use of CNN (Convolutional Neural Network)

Here, we have taken a single image and predicted the emotions score has it can be seen in the above image along with it we have also predicted the dominant emotion. The scores associated with the emotion would be further used in fusion with audio scores. Training Single Image and Resulting outcomes can be seen in the figure number 6



```
]: result = DeepFace.analyze(img1,actions=['emotion'])

   Action: emotion: 100%|██████████████████████████████████████████| 1/1 [00:00<00:00,  3.12it/s]

]: print(result)

   [{'emotion': {'angry': 0.22313743829727173, 'disgust': 5.724342956134393e-16, 'fear': 4.907389494768210 6e-08, 'happy': 99.77682
   828903198, 'sad': 4.0591835448466895e-08, 'surprise': 8.173253740473285e-10, 'neutral': 3.744853813714144e-05}, 'dominant_emoti
   on': 'happy', 'region': {'x': 82, 'y': 21, 'w': 126, 'h': 126}}]
```

Fig. 6. Training Single Image and Resulting outcomes.

B. Speech Emotion Recognition

For Speech Recognition we have downloaded the dataset related to it from Kaggle and the name of the dataset is EmoDB. Here, we have made use of three feature extraction techniques: MFCC (Mel-frequency cepstral coefficients), Mel- Spectrogram, LPC (Linear Predictive Coding) and two models: CNN - Convolutional Neural Network,

LSTM - long short-term memory networks.

The Accuracy that we got after training each model with corresponding feature extraction inputs are listed in the table

below IV-B:

Table IV-B : Feature extraction using different techniques

	CNN	LSTM
MFCC	66	88
LPC	37	40
Spectrogram	44	14
MFCC and Spectro	72	55

So, after training the models with different features extraction techniques we decided to go with MFCC feature extraction technique and LSTM Model has it has a accuracy of 88 percent. The below graph depicts the train accuracy vs the testing accuracy of our final model i.e LSTM using mfcc features extraction technique.

can be seen in the figure number 7

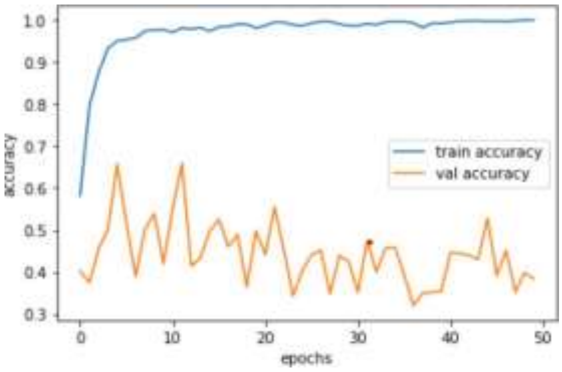
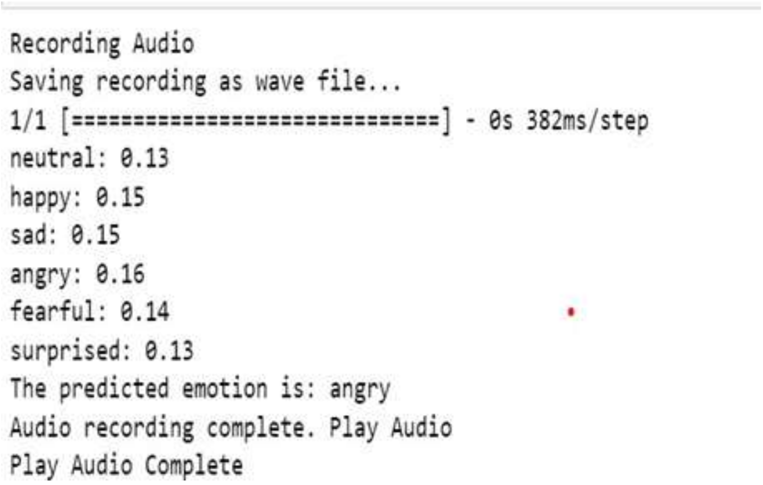


Fig. 7. Graph of Accuracy v/s Epochs.

Before fusing the audio and video part in our project, we have implemented the real time audio emotion recognition by taking live audio input from the audio port of our device and along with it obtained scored related to each emotion and predicted the dominant emotion in each audio. The sample Output can be seen in the figure number 8:



```
Recording Audio
Saving recording as wave file...
1/1 [==============================] - 0s 382ms/step
neutral: 0.13
happy: 0.15
sad: 0.15
angry: 0.16
fearful: 0.14
surprised: 0.13
The predicted emotion is: angry
Audio recording complete. Play Audio
Play Audio Complete
```

Fig. 8. Live Audio Input

C. *Fusion*

After predicting the emotions separately from real time audio, video and obtaining the dominant emotions along with scores related to it. Here, fusion of the emotions will be done. The dominant scores from the video will be taken and the dominant scores from the audio will be taken from each 1 second frame where there would be 20FPS in video and 20 coefficients in audio and that would be fused using weighted average graph method.

VI. RESULT

Hence, we can conclude that detecting emotions is crucial, especially in a world where people have already gone digital and like to socialize. Besides, certain deep learning models can allow us to create projects like this. This project aims to detect the emotions of people digitally and thereby take action or design solutions in the future. Our model detects emotions on a real-time basis. Once the code is in a running state, it opens the webcam of the system and whatever expression is being displayed (in terms of facial expression and speech) is detected.

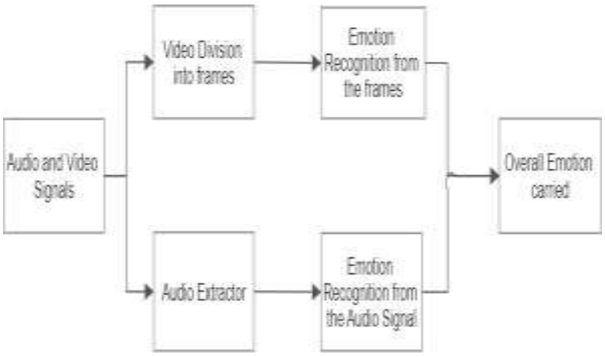

Audio and Video Signals
Video Division into frames
Emotion Recognition from the frames
Audio Extractor
Emotion Recognition from the Audio Signal
Overall Emotion carried

Fig. 9. Block Diagram of Fusion

Here is the final result that we obtained by taking realtime audio and video input:

CONCLUSION AND FUTURE WORK

In conclusion, emotion recognition based on speech and facial expression has gained significant attention in recent years due to its potential applications in various fields such as psychology, human-computer interaction, and marketing.

With the development of machine learning algorithms, researchers have achieved remarkable progress in recognizing emotions from speech and facial expression data.

Speech-based emotion recognition methods have been successfully used in various applications such as call center analysis, affective tutoring systems, and emotion-based music recommendation systems. Facial expression-based emotion recognition has also been used in many areas such as security systems, entertainment, and healthcare.

However, there are still some challenges that need to be addressed in the future work. One of the main challenges is to improve the accuracy of emotion recognition algorithms. Moreover, there is a need for more large-scale datasets that can capture a wide range of emotions and ethnicities. The use of multimodal approaches that combine speech and facial expression data is another area that requires further exploration. Additionally, the development of real-time emotion recognition systems that can work in real-world scenarios is also a significant area of research. In the future, emotion recognition based on speech and facial expression has the potential to bring significant benefits to society. It can help individuals better understand their emotions, assist in mental health diagnosis and treatment, and improve human-computer interaction.

REFERENCES

[1]. Linqin Cai, Jiangong Dong, and Min Wei. “Multi-modal emotion recognition from speech and facial expression based on deep learning”. In: *2020 Chinese Automation Congress (CAC)*. IEEE. 2020, pp. 5726–5729.

[2]. Ninad Mehendale. “Facial emotion recognition using convolutional neural networks (FERC)”. In: *SN Applied Sciences* 2.3 (2020), p. 446.

[3]. Danai Moschona. “An Affective Service based on Multi- Modal Emotion Recognition, using EEG enabled Emo- tion Tracking and Speech Emotion Recognition”. In: Nov. 2020, pp. 1–3. DOI: 10 . 1109 / ICCE - Asia49877 . 2020.9277291.

[4]. Gou Wei, Li Jian, and Sun Mo. “Multimodal (audio, facial and gesture) based emotion recognition challenge”. In: *2020 15th IEEE International Conference on Auto- matic Face and Gesture Recognition (FG 2020)*. IEEE. 2020, pp. 908–911.

[5]. Lin Xi et al. “2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG”. In: ().