EMOTION RECOGNITION ON VIDEO SPEECH

Prof. Tamal Dey, *Assistant Professor* and Riya Mandal, *MCA*

Dept. of Computer Applications, PES University, Bangalore, India - 560085

***Abstract*—To build a Machine Learning model that can process video speech signals and detect human emotions like Happiness, Anger, Sadness, Surprise, etc. The main processes involved are Preprocessing audio chunks from YouTube Videos, Extract the acoustic features and classify them discretely using a pre-trained model with the help of a pre-trained dataset. The main purpose of this project is to identify the emotions of the speaker in the video from their speeches.**

Keywords*—*Video, Emotion Recognition, Audio Chunks, MLP Classifier, Extract, Feature Extraction, Speaker

I. Introduction

A human recognition system is the need of modern tech dominated era where everything is automatic and customers are targeted through AI. It is very important for Computers to understand the human sentiments as people does so that in near future they can understand better about emotions and trained according to people's needs. The work in this field is ongoing and it will take about a decade before the human emotion recognition comes anywhere near accurate. However, there are many research paper and software addressing to this complex problem and someday the humans can understand us even better than other people. The idea behind this project is to calculate human emotions of videos like anger, sadness, etc. through machine learning algorithm. In doing so, YouTube videos comes handy to download and process it into audios and audio features such as its tune and wavelengths to detect the underlying emotions. Pre-trained dataset needs to be used for training the model to understand the emotions and features better.

II. Literature Survey

Emotion Recognition is the process of detecting human sentiments by the computer software through Machine Learning. In this era of AI and Automotive Industry detection of emotions plays a vital role to understand the people and customers. Human emotions such as Happiness, Anger, Sadness, Excitement, etc... are very complex sentiments and differs from person to person and situation wise as well. People have different perspective to respond to same situation. However, the new progress in Machine Learning Algorithms has shown promising results in this area. So far, the progress is very ground level and the works are going on to analyze human emotion better but perfection is a long road and the research works continues.

Economic Feasibility : This project does not need any costly computer and hence it is affordable. All the libraries used are open source and hence no license needs to be purchased.

Operational Feasibility : The website is user-friendly so that users can easily use this website without any confusion and users can select the audio from their local files to predict emotions.

There are many papers written in this field that has been referenced for this project. These related works are:

Title : HUMAN EMOTION RECOGNITION FROM AUDIO AND VIDEO SIGNALS

Author: Sai Nikhil Chennoor, B. R. K. Madhur, Moujiz Ali, Dr. T. Kishore Kumar

Summary: The paper deals with different audio and video dataset such as Berlin Emo database and JAFEE database. The paper demonstrates a model for emotion recognition from audio and video signals that can be deployed in devices with nominal capabilities.

Implementation strategy: The video signals are transformed to frames and emotion on these frames are recognized and makes the overall emotion of video signals. The LBP feature extraction and SVM classification model is used for video frames. Similarly, Audios are extracted from videos and SVM model is used on features like pitch and entropy to calculate the emotion. At the combining stage a threshold value is given from 0 to 10 and if the difference between first 2 maximum frame counts is greater than threshold then emotion from video signal is counted else emotion from audio signal is counted.

Result: Three hundred best video clips are selected, and 225 (75%) video clips are taken for training, and 75(25%) video clips are taken for testing giving 74.66% accuracy. Same files were taken for emotion detection audio signals giving an accuracy rate of 69.33%. The complete model accuracy is 77.33%. [5]

Title : SPEECH EMOTION RECOGNITION WITH DEEP LEARNING

Author: Pavol Harár, Radim Burget and Malay Kishore Dutta

Summary: This paper deals with Speech Emotion Recognition (SER) using Deep Neural Network (DNN) architecture with convolutional, pooling and fully connected layers. German Corpus dataset has been used with 3 class subsets angry, neutral and sad.

Implementation strategy: The Stochastic Gradient Descent algorithm is used on DNN architecture. The input data were presented in batches of size 21 in multiple epochs. The maximum value from predicted probabilities is taken to denote the predicted class. The average probability is computed for all

segments belonging to the particular file and used it to denote the final predicted class.

Result: The experiment resulted in 79.14% validation accuracy and DNN achieved 96.97% accuracy on testing files. The average confidence of file prediction was 69.55%. [6]

Title : "EMOTION RECOGNITION IN SPEECH USING CROSS-MODAL TRANSFER IN THE WILD"

Author: Samuel Albanie, Arsha Nagrani, Andrea Vedaldi, Andrew Zisserman

Summary: This paper is based on a hypothesis that the emotional content of speech correlates with the facial expression of the speaker. The VoxCeleb Dataset is used.

Implementation strategy: Teacher-student method is used where a strong teacher network that performs emotion recognition from face images training teaches student model, which is tasked with performing emotion recognition from voices. The CNN network is trained for 50 epochs.

Result: The student achieves a mean ROC AUC of 0.69 over the teacher-predicted emotions present in the unheard identities (these include all emotions except disgust, fear and contempt) and a mean ROC AUC of 0.71 on validation set of heard identities on the same emotions. [7]

III. METHODOLOGY

Data Preparation

Download YouTube Videos: Download a few videos from You Tube with clear audios and speech of English-speaking speakers.
Convert to .wav format: Convert the downloaded videos into .wav audio format.
Clip Audio: The audios need to be clipped at the start and end to avoid unnecessary noisy part of the videos.
Extract Audio Chunks: The toughest part of data preparation of dividing a long audio file in smaller audio chunks of few seconds duration. To chunking audios manually one by one is a very long tiring process. So, to make the process easier the audios are split on the basis of silence period and threshold using python libraries.
Cleaning and masking Audio Chunks: The chunks are further cleaned and saved in different folder where the trained model will be applied.

Data Visualization

Play Audio: The audios can be played in Jupyter notebook to check the quality and tone.
Plotting Graphs: The graphs are plotted for the audios to see in depth qualities of sounds and checks on features such as MFCC, MEL on which the emotions of speaker are to be classified.

Training RAVDESS dataset:

The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) is pretrained dataset containing 1440 audio files which is used to train the model for Emotion recognition in this project.
Download RAVDESS Dataset: The RAVDESS dataset is available online for the sole purpose of emotion recognition.
Extract Features: Extract features like MFCC, MEL, CHROMA from audios.
Select Emotions: The emotions on which features to be observed and applied to Model like Happy, Angry, etc.
Split into Train and Test data: Split the whole dataset into train and test.
Apply MLP model: The Multi-Layer Perceptron classifier is used for classification of emotions.
Predict Test Set: The test set of data is predicted.
Check Accuracy: Accuracy of the model is important deciding factor. More the accuracy rate better the model works on data.

Save the model: Save the trained model into local file for future use.

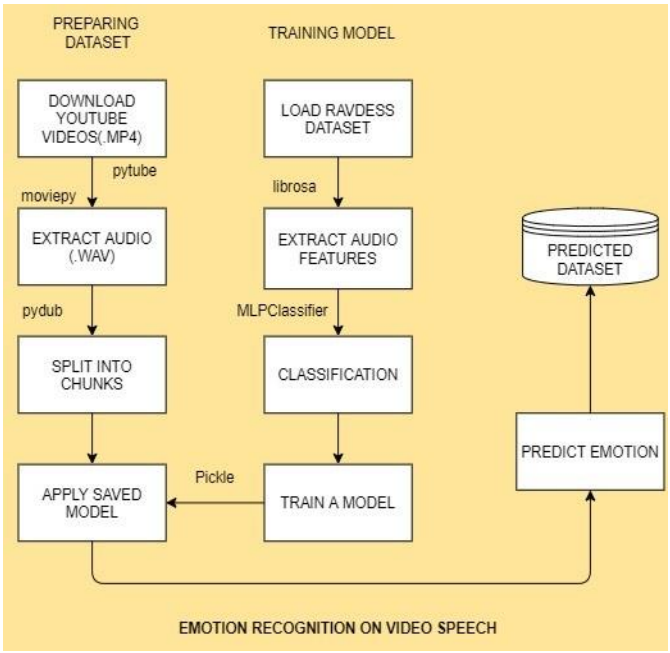

Fig. 1. A Flow Diagram for the project Emotion Recognition on Video Speech

Apply Model into Dataset created from YouTube

Load Saved Model: Load the saved model file into Jupyter Notebook,
Extract Features: Extract features Such as MFCC, MEL, CHROMA of the audios.
Deploy Model: Apply the model into the self-created dataset to predict emotion of the audio chunks.
Save Predicted Emotion into CSV File: The predicted emotions are to be saved for future use like visualization.

Predicted Emotion Visualization

The predicted emotions are visualized using bar graph to show the emotion distribution in each video and overall dataset created from different videos. It also helps to show the dominated Emotion in a video.

IV. DESIGN AND MODELING

MLP Model:

Since RAVDESS is a labelled dataset, A supervised learning is implemented in the project. Emotions were classified into 4 categories i.e., Happy, Anger, Disgust and Calm. For such big dataset Artificial Networks need to be used and hence the MLP classification model is used as it gives maximum accuracy for emotion prediction. The Multi-layer Perceptron Classifier is initialized with:

model=MLPClassifier(alpha=0.01,batch_size=256,epsilon=1e-08, hidden_layer_sizes=(300,), learning_rate='adaptive', max_iter=500)

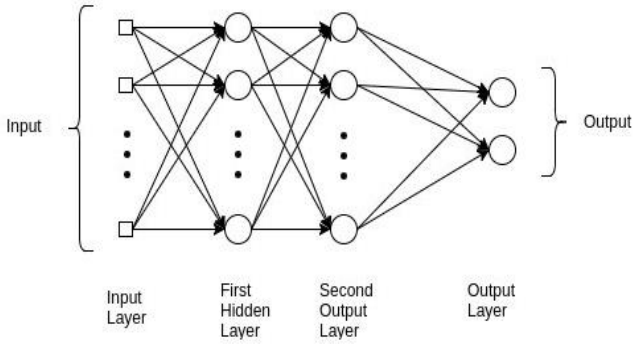

Fig. 2. A diagram of MLP model network

The above fig 2 shows the internal working of an MLP network. The audio signal features like MFCC, MEL Chroma values are put into the Input Layer of the model. Here hidden layer is only one with 300 neurons (each node is a neuron) and that processes the feature values. The output layer gives the predicted emotion.

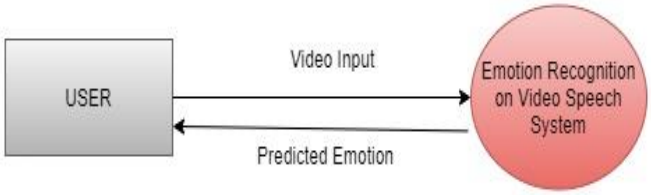

Fig. 3. A context diagram of the emotion recognition on video speech

The above diagram, fig 3 shows the working of this project. A user gives input in the form of audio chunks and the system, in this case a User Interface gives the predicted emotion for given audio signals.

V. RESULT

The predicted dataset is saved into csv files and analyzed based on emotion distribution.

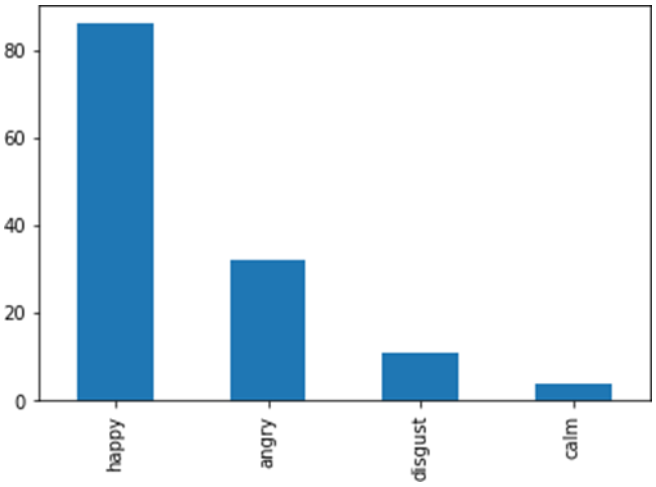

Fig. 4. The emotion distribution between audio chunks of a single video file

After complete prediction is done in the created dataset the analysis is done for every audio chunk in a video file and put in the form of a bar graph (fig 4) to showcase the findings. The emotions that are taken for prediction are “happy”, “anger”, “disgust” and “calm” It is clear that since the videos are mostly public speeches the emotion that dominates all is “Happy” followed by “Anger”.

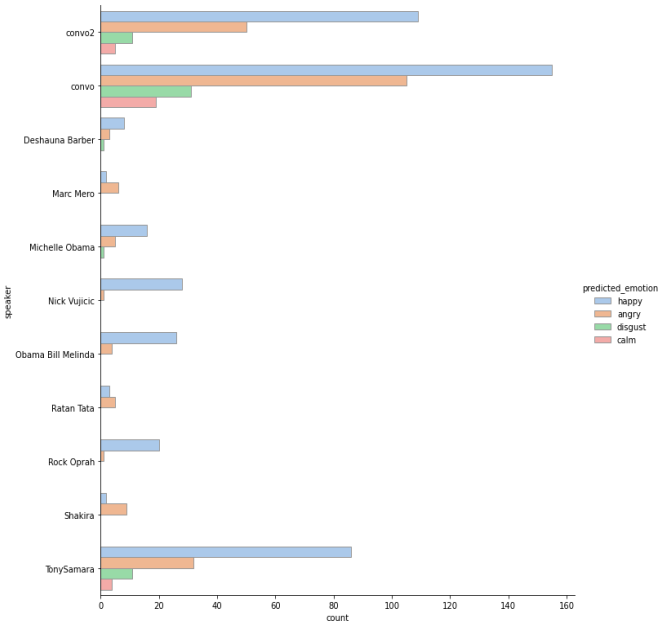

Fig. 5. A collective emotion distribution of all files and all audio Chunks.

The above graph (fig 5) is representation of every video file that is taken for this project and depicted emotion distribution for every one of them just like fig 4. The y- axis represents the name of the speakers in every video that has been downloaded from YouTube.

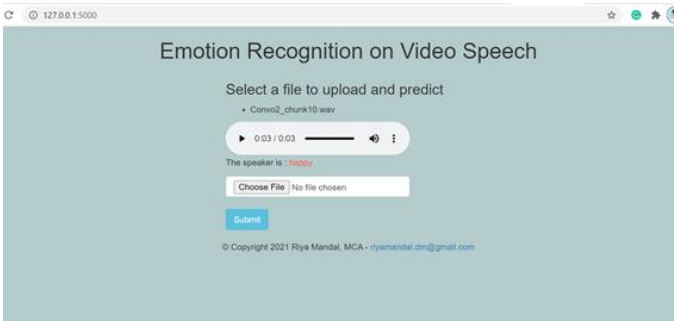

Fig. 6. User Interface to select the audio file and predict emotion.

The project is shown more clearly with the help of a user interface (fig 6) that enables the users to upload a audio file in .wav format and the website will predict the emotion of the speaker. The User interface is very simple and very easy to use.

VI. CONCLUSION

The accuracy rate of 81.82% is achieved with MLP model and 4 emotions i.e... Happy, Calm, Angry and Disgust. The accuracy rate differs for the combination of different Emotions. The emotion of the speaker changes in audio chunks over a period of time in video. Emotions like Happy and Angry is dominated in most videos.

Since the dataset are made from Random YouTube videos, finding emotion variation in single video with clear audio is difficult. A better trained model is required for better accuracy. The model takes tones and pitch and don't identify words and context of speakers. So, the emotion detected may vary from a person's point of views. More works need to be done in the field of emotion prediction.

REFERENCES

[1] [Online].Available:https://www.geeksforgeeks.org/pytube-python- library-download-youtube-videos/

[2][Online].Available:https://www.analyticsinsight.net/speech-emotion- recognition-ser-through-machine-learning/

[3] [Online]. Available: https://www.codespeedy.com/extract-audio-from- video-using-python/

[4] S. Bennett, Object Oriented Systems Analysis and Design Using, 4th ed., McGraw Hill,, 2010.

[5] S. N. Chennoor, B. R. K. Madhur, D. T. K. Kumar, and M. Ali, "HUMAN EMOTION RECOGNITION FROM AUDIO AND VIDEO SIGNALS."

[6] Pavol Harár, Radim Burget and Malay Kishore Dutta "SPEECH EMOTION RECOGNITION WITH DEEP LEARNING"

[7] Samuel Albanie, Arsha Nagrani, Andrea Vedaldi, Andrew Zisserman "EMOTION RECOGNITION IN SPEECH USING CROSS-MODAL TRANSFER IN THE WILD"