Stress and Anxiety Detection through Speech Recognition and Facial Cues using Deep Neural Network

Sharmila Chidaravalli[1], Namratha Jayadev[2], Divyashree P[3], Ghanavi Yadav A[4], Prajwal B[5]
[1,2,3,4,5] *Dept. of Information Science & Engg., Global Academy of Technology, Bangalore, India*

Abstract **- Stress is a feeling of emotional tension and anxiousness. It can have an influence on each individual's mental health. On the other hand, anxiety is a common reaction to stress which makes one fearful thus leads to panic attacks. Stress and anxiety can lead to unreasonable complications with an individual's personal life. These mental issues can cause mental instability which has to be treated in a right manner. This paper analyses how we use vocal/audio dataset and video visuals i.e. facial expressions to detect stress and anxiety in an individual. Here we have developed a model where stress and anxiety is detected using deep neural network. We use these actors audio/vocal datasets from Kaggle where the audio consists of 7 emotions i.e., anger, surprised, sadness, neutral, disgust, fear and joy. Later the audio datasets are used to train and test few of the classification models like Convolution Neural Network (CNN). Then the audio which is collected will be pre-processed through acoustic feature extraction, accordingly the audio is classified through CNN which provides the accuracy based on those 7 emotions. CNN is also applied to analyze visuals where it derives a relationship between pixels by determining features of an image using input data. Input image is passed through convolution layers with filters like kernel to produce the outcome of facial expressions. By this we can predict if the person has stress or anxiety.**

Index Terms **- Convolutional Neural Network, Emotion Classification, Stress Detection, MFCC (Mel frequency cepstral coefficients), Chroma.**

I.INTRODUCTION

Stress is one of the most common problems that occur in daily life of every individual. Stress and anxiety are the emotional states that usually affect mental and physical health of humans. Several researches have proven that these states are connected to one another. Monitoring the levels of stress or anxiety can help in preventing major disorders such as generalized anxiety disorder (GAD). There are different ways in which human body reacts to the stress. Stress can be divided into two types, short-term stress and long-term stress. Short-term stress is usually due to changes in the situation and it is for a short period of time. Long-term stress is due to major problems in life, which is dangerous. The emotional stress can be identified by facial expressions and voice modulation. Our daily life involves a lot of communication with other people, facial expressions is also a means to communicate. There are mainly five emotions that are universally accepted. They are happy, sad, anger, disgust and fear. On a daily basis human recognize the emotions by analyzing the features displayed on the face. For example, happiness is connected with a smile on the face. Emotions can also be recognized by the voice modulation which requires human-computer interaction. Speech is given as an input to the machine for speech analysis. The input is translated into text format which is known as Speech Recognition System or Speech to Text. The speech recognition system analyses an individual speech in order to determine the emotion and produces accurate result. The speech signal which is extracted is trained by DNN model. Finally, the output obtained will be compared with connected and continuous speech. Images are captured through a webcam for video analysis. All the necessary features will be extracted and then the model predicts whether the individual is under stress or not.

Human behavior depends on the way humans act and interact with others. Analyzing human behavior is a very important practice mainly in psychotherapy. Behavior can be analyzed by observing the way in which emotion changes during the conversation. Here, we implement deep learning in order to analyze the emotional state. We have determined the relationship between emotion and behavior, further use emotions to classify the behavior of an individual. In our system, we take the input speech and images to determine

signals/facial cues and then predict whether the individual is under stress or not.

II. RELATED WORK

In the work done by Jose Almeida and Fatima Rodrigues [1] they proposed a video system to detect Stress. The video images are captured via computer's webcam. Later the captured image are cropped and resized to the desired pixels using the Haar like technique. Emotion classifier on the resized facial expression in applied by implementing Convolutional neural Network (CNN). They made use of the seven facial expression to classify whether a person was undergoing Stress or not. They have used two different dataset for this approach CK+ dataset and KDEF dataset. The neural network which they proposed was VGG16, VGG19 and InceptionResNetV2. After training the models they were compared to know which model yields the highest accuracy. The VGG16 model obtained a highest accuracy of 92.1% compared to the other two models.

Muhammad Arif, Ashjan Basri, Ghufran Melibari, Taghreed Sindi, Nada Alghamdi, Nada Altalhi and Maryam Arif [7] have bandied about bracket of anxiety diseases using machine learning styles. Machine learning algorithms can be used to classify the presence/ absence of a particular anxiety complaint, vaticination of threat situations, or vaticination of response situations of treatment. Data can be collected from different coffers including demographic data, health records, medical history, different measuring scales, etc. A large set of features can be attained from the data and important features may be named grounded on an applicable point selection algorithm. These features correspond of training and testing data set for the classifier. Grounded on the training data set, selection, and parameter tuning of a good classifier grounded on performance criteria is the last step. Different features of audio speech like pitch, speaking rate, articulation, specific spectral and timing parcels can give indication about the depressed person (172). These features of the audio signal can also be used as anxiety predictors in the speech Hence, analysis of audio conversations/ drooling on social media can be anatomized to know the emotional status of the person and these suggestions may be helpful in the early discovery of anxiety. Most of the researchers have used SVM, and random forest classifiers for the detection of different types of anxiety disorders. Results were given according to different type of disorders in which SVM and random forest method gave 99% and 92% accuracy respectively.

Dr.S.Vaikole, S.Mulajkar, A.More, P.Jayaswal, and S.Dhas [3] proposed an algorithm that first extracts Mel-filter bank coefficients using a pre-processed speech data and then predicts the stress output using CNN. The audio signal is passed to speech preprocessing and then forwarded to feature extraction module. All the necessary speech features are extracted and are passed to a deep-learning based stress detection model. The CNN model determines the user's stress state by a decision process. The proposed system uses Ravdess database. Total of 1440 Speech utterances of twelve male and female speakers were taken. Labels were used for training the model using one-hot-encoding approach. The accuracy was classified into pitch rate and MFCC. The proposed model consists of eight CNN layers and fully connected layers. These layers capture the necessary information of extracted features and then calculate the frame-level output each time. The output of frame-level is converted into a sentence-level feature. The features extracted from layers are of two types that is average value of output sequence and last frame-level output. The accuracy of stress detection system using pitch rate was 52% and using MFCC was 94.33%. He further concluded that by using signal raw energy operator stressed emotions are detected with improved accuracy.

Faizan Ahmad, Aaiman Najam and Zeeshan Ahmed [4] have discussed that the image/video of a person by using facial recognition provides accuracy and speed in the biometrics research. They have used rich face datasets in terms of subjects, light, pose, rays and emotions for classification. They have implemented AdaBoost classifier along with Haar, Local Binary Pattern (LBP) to enhance the features whereas Support Vector Machine (SVM) which is used with Histogram of Oriented Gradients (HOG) these algorithms are used for the facial detection evaluation. The image is given us the input which generates a set of features Haar will evaluate the features and uses the AdaBoost algorithm for the extraction. The LBP helps in labeling the pixels of an image to 3 by 3 neighborhoods including the central pixel value then the operator divides the image into 'n' overlapping regions. SVM

with HOG is used for face detection which reduces the sensitivity to facial position, SVM returns the binary value. SVM is trained to capture the dissimilarity between two images the obtained results from the algorithms are: AdaBoost which consists of Haar and LBP has the accuracy of 96.70 % and 89.30 %. SVM which consists of HOG has the accuracy of 90.88%.

Reshma Radheshamjee Baheti and Supriya Kinariwala [9] proposed another method to detect Stress using Machine Learning technique based on the tweets on the social media. They have used Twitter Sentiment Dataset for this purpose. The texts undergoes a pre-processing stage which includes removal of special characters, removing extra spaces form the sentences, removing URLs and removing the words which does not help in the classification stage. This pre-processed text is subjected to Word Sense Disambiguity (WSD) technique where each word is tagged to a particular English Parts of Speech. They have implemented TensiStrength which helps in identifying the sentimental strength of the text. For the classification and prediction they implemented SVM and Naïve Bayes. SVM gave an accuracy of 67%.

III.METHODLOGY

A. Dataset Collection

The actor based speech database is comprised of 2768 files for audio data sets. On emotional validity, strength, and genuineness, each file was scored 10 times. There were 24 individuals that were characterized by an un-trained adult study candidates belonging to North America were given scores. High emotional validity levels, reliability of interrater, and reliability of test-retest interrater were recorded. In the database, there are 24 trained actors (12 male, 12 female), in a North American neutral voice, clearly expressing two linguistically related phrases. Speech includes expressions of neutral, happy, sad, angry, fear, disgust surprise and calm. At two emotional intensity ratios, (strong and normal), each expression is generated with an additional neutral expression. There are three mode formats available for all conditions: audio-only (16bit, 48kHz.wav).

B. Audio Impementation

Speech recognition is the way of converting acoustics (speech of a person) into textual form. The google API called Speech Recognition which allows us to convert speech into textual for further processing. Firstly, we internally see the input physical audio which will get converted into electric signals. The electric signals of our speech signal then gets converted into digitized form with an analog-to-digital converter. Any type of sound created by humans is defined by their vocal tract shape, including tongue, teeth, lips, etc. The envelope of the time power spectrum of the audio signal is representative of the vocal tract and MFCC, defined as the coefficients that make up the Mel-frequency cepstrum and correctly represent this envelope. Chroma relates to the twelve different kinds of pitch classes and tuning approximated to the equal tempered scale. It basically computes melodic and harmonic characteristics of speech or an audio signal.

The only audio feature to train our CNN model, the MFCC and Chroma features are considered the basic approach. The MFCC coefficients were only used for their ability to reproduce the amplitude spectrum of the audio wave in a compact vector form. The discrete Fourier transform is implemented, then the logarithm of the amplitude spectrum is taken into account. After a certain amount of frequency 'Mel' reduction, the spectrum of amplitude is then normalized. For a significant re-construction of the sound wave that can be distinguished by the human auditory process, this technique is performed to empathize the frequency to a more realistic type. For each speech file, some features were extracted. Features were produced and along with it converting each speech file to a time series of floating points Then MFCC sequence was created from the time series.

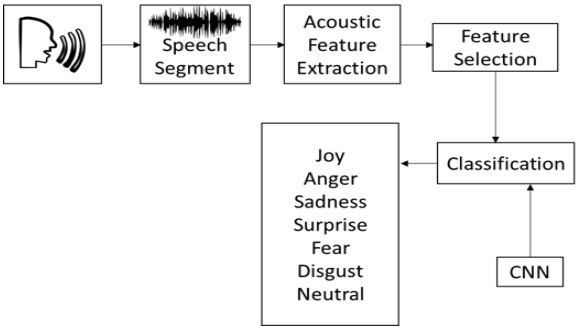

Fig.1: Audio System Overview

If the input given is a size < set of training samples > x n x 1 on which we executed a one-dimensional CNN round as the activation function ReLu and 2 x 2 is the max-pooling function. Relu layer is added to the function to represent the hidden units. The last activation layer is used as the SoftMax layer which

calculates relative probabilities. Then at the end the fully connected layer is used where the classification happens. Pooling allows the CNN model to focus only on the main characteristics of each of the data components, not segregating them by their position. The output of the pooling layer is flattened and this flattened matrix is fed into the fully connected layer.

C. Video Implementatiom

For the video recognition purpose, we are implementing CNN model. The image is captured using the computer's webcam. After the image is captured Haar classifier called Viola-Jones is implemented and used. Further the image is subjected to sharpening and restoration to create a better image. Later the image is resized to the area of interest. After the image pre-processing is done, important facial features are extracted using LBP algorithm. In LBP algorithm each pixel is compared amongst its adjacent eight neighbors in a 3x3 matrix. The negative values are encoded with 0 and the remaining with 1. After the feature extraction process, it is forwarded to the CNN classification model where it classifies the images to one of the seven emotions. CNN is better in classifying image since it is able to capture special features.

The proposed model consists of four convolutional layers, max pooling layers and fully connected layers. In convolutional layer after the computer reads the input image in the form of pixels. We take a small patches of the image this is called the features. We send this features and it gets a lot better at seeing similarities. It predicts the class probabilities for each of the facial features by applying a filter that scans the whole image. In ReLU layer the negative values are converted to 0, this is done to avoid the values from summing up to zeros. It activates the node only if the input value is above a certain value while if the input is below zero the output will be zero and all the negative values are removed from the matrix.

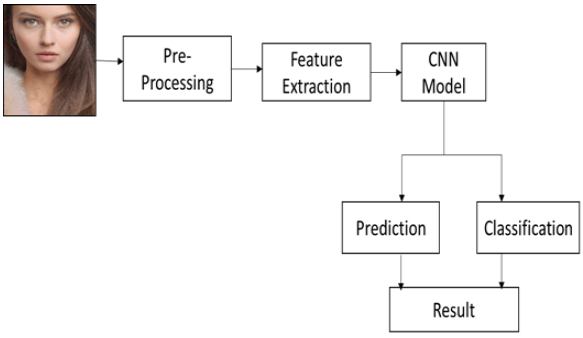

Fig.2: Video System Overview

In pooling layer we reduce the image size. It scales down the amount of information the Convolutional layer generated for each feature and maintains the most essential information. We pick a window size then mention the required stride, then walk the window across the filtered images. In fully connected layer the output generated by the previous layer is converted to a single vector that is used as an input for the next layer. It applies the weight over the input generated by the feature analysis to predict the accurate label. These layer are repeated until we get a 2x2 matrix. At the end of fully connected layer the output is determined under which emotion is the face classified.

IV.RESULT

The findings attained from the evaluation process indicate the efficacy of the model on the dataset relative to the baselines and the state of the art. It shows the precision, recall and F1 score values that were attained for each of the emotional groups. These findings suggest that recall and accuracy are kind of balanced, enabling us to achieve a 0.76 F1 score for the class. The slight shift in F1 highlights the robustness of the CNN model, which manages 76.08 percent accuracy effectively. The accuracy of the video is 80% as shown in Fig. 3.

V.CONCLUSION

This work presents a deduced model that takes audio and video as an input and identifies whether the user is under Stress and Anxiety. In this paper we have proposed a simple system to carry out the above mentioned functions. We have extracted the MFCC, MEL and Chromogram features from the audio files used throughout training to acquire such results, and implemented CNN model for extracting facial cues. We trained our neural network on the above representations of input data to correctly figure out the probability of distribution of annotation sections employing 1-Dimensional CNN, max-pooling and Dense Layers. The result gained can only be worth it as a starting point for further expansions, updates, and enhancements.

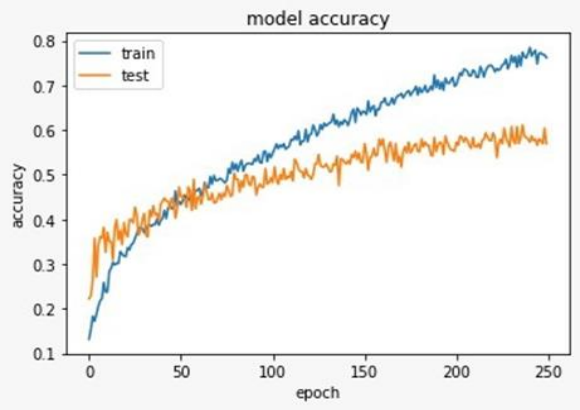

Fig 3: CNN Model Epoch vs Accuracy

REFERENCE

[1] Jose Almeida and Fatima Rodrigues, “facial Expression Recognition System for Stress Detection with Deep Learning”, 2021.

[2] Arushi Roberto Dilloa and Ai Ni Teoh, “Real-time Stress Detection and Voice Analysis: An Integrated VR-based Game for Training Public Speaking Skills”, 2021.

[3] Dr.S.Vaikole, S.Mulajkar, A.More, P.Jayaswal, "Stress detection through speech analysis using machine learning", May 2020.

[4] Faizan Ahmad, Aaima Najam Zeeshan Ahmed, "Image based face detection and recognition", 2020.

[5] Russel Li, Zhandong Liu, “Stress Detection Using Deep Neural Networks”, 2020.

[6] Zarinati Hosseinzadeh-Shanjanji, Khadijeh Hajimini, Bahram Rostami Shokoufeh Ramezani, Mohsen Dadshi, “Stress Anxiety Levels Among Healthcare Staff During COVID-19 Epidemic”, 2020.

[7] Muhammad Arif, Maryam Arif, Asjhan Basri- "Classification of Anxiety Disorders using Machine Learning Methods", 2020.

[8] Muhammad Arif, Maryam Arif, Asjhan Basri- "Classification of Anxiety Disorders using Machine Learning Methods", 2020.

[9] Reshma Radheshamjee Baheti, Supriya Kinariwala, “Detection and Analysis of Stress Using Machine Learning Techniques”, October 2019.

[10] Kevin Tomba, Joen Dumoulin, Elena Mugellini, Omar Abou Khaled and Salah Hawila, “Stress Detection Through Speech Analysis”, 2018.