

# Stress and Anxiety Detection through Speech Recognition Using Deep Neural Network

Divyashree P<sup>1</sup>, Ghanavi Yadav A<sup>2</sup>, Namratha Jayadev<sup>3</sup>, Prajwal B<sup>4</sup>, Sharmila Chidaravalli<sup>5</sup>

<sup>1,2,3,4</sup>Students, Global Academy of Technology, Bangalore

<sup>5</sup>Assistant Professor, Global Academy of Technology, Bangalore

**Abstract**—Stress is a feeling of emotional tension. It can have an influence on our mental health and for the people around us. While anxiety is a natural reaction to stress which can be fearful this can lead to panic attacks. These mental issues have to be addressed by everyone. This paper explains how we are using vocal/audio dataset to detect stress and anxiety in a person. We have developed a stress and anxiety detection model using deep neural network. Here audio datasets is considered from Kaggle in which the audio consists of 7 emotions i.e., joy, fear, disgust, neutral, sadness, surprised and anger. These audio datasets are used to train and test classification models like CNN. Then the audio is pre-processed through acoustic feature extraction, classified through CNN which provides the accuracy based on those 7 emotions. By this we can predict if the person is stressed or has anxiety

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

## I. INTRODUCTION

The speech recognition aims to determine the emotional state of an individual by using his/her voice. Speech is a means of communication to express one's thoughts and feelings. It is one of the fast and best ways to communicate. Speech recognition is most beneficial in applications that require human-computer interaction such as speech synthesis and customer service. Recognizing the emotional state of an individual using speech signals can be difficult for several reasons. The speech processing applications has a great influence in our life on commercial applications like Text-to-Speech synthesis, Speech recognition and verification. Speech is given as an input to the machine that accepts this command. 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.

Emotions act as an important part in each day of human interactions. Emotions can be happy, fear, sad, anger and disgust. It is necessary to our rational and intelligent decisions. It helps us to communicate and determine the feelings of other people by communicating our feelings and responding to others. Emotion plays an important role in shaping our behavior. It displays the information about mental condition of an individual. The speech signals can be easily detected by using a microphone. This feature is very useful for the users and it also helps to maintain, build a large database for stress detection system.

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 used emotions to classify the behavior of an individual. In our system, we take the input speech to determine speech signals and then predict whether the individual is under stress or not.

## II. RELATED WORK

Recent researches for detecting Stress and Anxiety has been extremely prominent. Sometimes the responses from Stress allows the body to overcome tough situation and prepare for treats but in contrast it can damage one's health too. There has been a lot of different methodologies for detection of Stress and

Anxiety through physical test, questionnaires that primarily rely on user input data which sometimes may not be accurate or user may find it difficult to answer some question if it is personal and sometimes measured through the speech modulation and frequency through which one person says his thoughts to others.

In the work done by Maghilnan S, Rajesh Kumar M [1] work he implemented Sentiment Analysis by Speech Data by proposing four steps 1) pre-processing which includes VAD. The input signal is given as an input to VAD which identifies and segregate the voice from the signal. The voices are then stored as chunks in the database. 2) Speech Recognition System. Here the words in the language spoken by the humans are converted to machine readable format which is processed further. The tools used for speech recognition are Bing speech, Google Speech Recognition. 3) Speaker Recognition System where the chunks are recognized and each chunks are identified and given the Speaker Id it helps in identifying whether the chunks are from the same speaker or different. The system then matches the Speaker Id with the system generated text. For feature extraction they have used Mel Frequency Cepstrum Coefficient (MFCC) and for feature matching they have used Dynamic Time Wrapping(DTW) 4) Sentiment Analysis they have implemented different algorithm such as Naïve Bayes, Linear SVM and VANDER and a comparison is made to find the efficient algorithm. The accuracy for Naïve Bayes was obtained as 72.8%, Linear SVM 86.4% and VANDER as 95.2%.

Kevin Tomba , Joel Dumonin , Omar Abou Khaled , Satish Hawalia [2] have discussed about multimodal stress classification system and utilized the audio/video data to investigate complete number of audio and video features with various fusion techniques and temporal backgrounds for classification purposes. They showed that Teager energy cepstral coefficients (TECC) surpassed standard baseline characteristics in the audio modality, while vector modelling depending on MFCC characteristics attained the best precision, while on the other hand, polynomial parameterization of face image characteristics produced the desired output across all systems and exceeded the best baseline system. MFCCs are used as features in the extraction model to extract the features. Three

different datasets were used the Bernil emotional database RAVDESS database and Keio University Japanese Emotional speech database. EmoDb and RAVDESS database were implemented using SVM and KeioSD dataset using ANN. Both SVM and ANN were optimized with the help of Scikit-learn library method. This method was used in finding the best combination of values to give the best result for a set of features. SVM and neural networks were used in the classification. Both algorithms showed best results, with ANNs having slightly better scores than SVMs. The obtained results performed good classification and determined if there is stress or not.

Anderson R. Avila, Shruti R. Kshirsagar, Abhishek Tiwari, Daniel Lafond, Douglas O'Shaughnessy and Tiago H. Falk [3] he used a CNN, SVM and DNN learning techniques and evaluated which model yields the highest accuracy. This experiment was performed using Speech Under Stimulated and Actual Stress (SUSAS) dataset. They proposed the use of modulation spectral features (MSF) as an input to CNN and adopted OpenSMILE features and evaluated it with SVM and DNN. In order to extract modulation spectral features the speech signal is first normalized to -26dBov and eliminating unwanted speech signals. Then they have filtered the signals using k modulation filters later the frequencies from the filter center are equally spaced from 4 to 128Hz. Finally five feature set are extracted. The results showed that the proposed MSF combined with CNN outperformed the other two learning methods SVM and DNN and gave an overall accuracy of 72% while DNN method achieved 62% accuracy and SVM produced 61% accuracy.

Dr.S.Vaikole, S. Mulajkar, A. More, P. Jayaswal, S. Dhas [4] 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.

Arushi, Roberto Dillon, Ai Ni Teoh [5] proposed a VR-based stress detection model where the speaker's voice is analyzed on real-time basis where virtually the speaker's speaking skills to the audience will be improved by real-insights from the game which provides the support/feedback. They have taken the dataset Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS). They have constructed 3 classifiers models to extract the voice features Amplitude Envelope (AE), Root-Mean-Square (RMS) and Mel-Frequency Cepstral Coefficients (MFCCs). Using Random Forest, KNN & SVM training and testing of data is done. Machine learning Algorithms like Gaussian Mixture Model (GMM), Hidden Markov Model (HMM), Artificial Neural Networks (ANN) and Deep Neural Network (DNN). VR based stress detection model includes virtual environment, behavior of virtual audience, machine learning model development, feature selection, training and testing of model development. In this model they have kept 70% of data for training and 30% for testing the actor's voice dataset. The final results shows that random forest accuracy is 82%, KNN accuracy is 72% and SVM accuracy 57%, 5% and 24% accuracy has increased to detect the stress with features that includes RMS, AE and MFCC.

### III.METHODOLOGY

#### A. Dataset Collection

The actor based speech database is comprised of 2768 files. On emotional validity, strength, and genuineness, each file was scored 10 times. There were 24 individuals that were characterized by an untrained 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. Speech Recognition

Speech recognition is the way of converting acoustics (speech of a person) into textual form. This is widely used in virtual assistants like Rebecca, Siri, Alexa, etc. The google API called Speech Recognition which allows us to convert speech into textual for further processing but while using the Speech Recognition API, translating big or long audio files into text, it may give error messages because it is not that strong for large files of audio. 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. Then, the digitized model can be used to transcribe the speech into textual form.

#### C. Feature Extraction

Acoustic Features: In general, the more precise and very basic features of audio to recognize affect are considered to be duration, MFCC, energy and pitch. This has been supported by a many research work and found it to be the most correct acoustic features to emotions are duration and energy, while all the other features are of medium relevance.

Mel-Frequency Cepstral Coefficients (MFCC) depending on a linear cosine transformation (CT) of a log power spectrum performed on a non-linear Mel frequency scale, it is known as the spectrum of short-term control of an audio or sound. 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. Options are considered for the lower dimensions of the 1<sup>st</sup> thirteen MFCC coefficients as they represent

the spectra envelope. And its spectral data is indicated by the higher dimensions which are discarded. Envelopes are necessary for different phonemes to display the difference, so we can find phonemes through MFCC. Chroma: It is also called as 'Chromagram', 'Pitch class profiles', 'Chroma features', that 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. It is consisting of 2 features:

- Chroma Vector: It has twelve element expression of spectral energy.
- Chroma deviation: It is the twelve Chroma parameters standard deviation.

#### D. Convolutional Neural Network

The deep learning model depending upon the convolutionary neural network (CNN) is used and its dense layers have been used. As 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. As mentioned in, the speech file is split into frames, using a fixed window size.

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.

If the input given is a size  $< \text{set of training samples} > \times n \times 1$  on which we executed a one-dimensional CNN round as the activation function ReLu and  $2 \times 2$  is the max-pooling function. ReLu as  $g(z) = \max\{0, z\}$ , and it gets a large value in the case of activation by adding this 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.

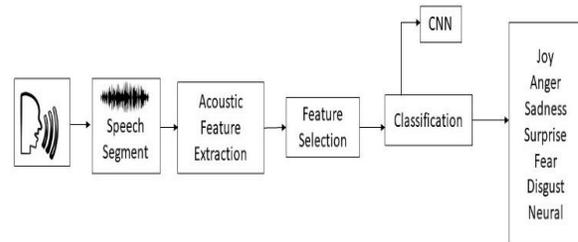


Fig 1: System Overview

#### 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.

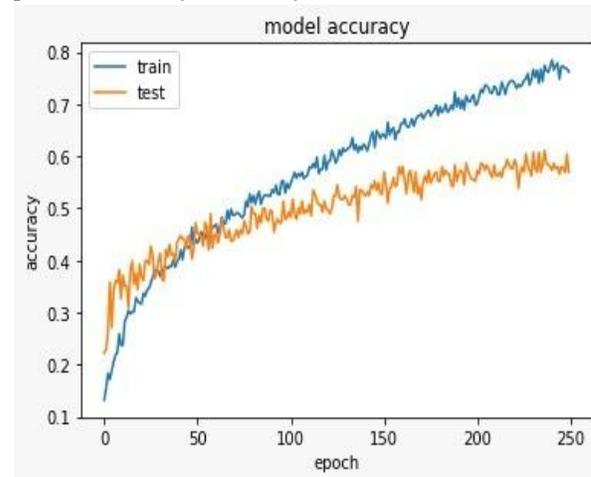


Fig 2: CNN Model Epoch vs Accuracy

#### V.CONCLUSION

This work presents a deduced model that takes audio 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 abovementioned functions. We have extracted the MFCC, MEL and

Chromogram features from the audio files used throughout training to acquire such results. 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.

Healthcare Staff During COVID-19 Epidemic”, 2020.

- [10] Arushi, Roberto Dilloa and Ai Ni Teoh, "Real-time stress detection model and voice analysis: An integrated VR-based game for training public speaking skills", 2021.

#### REFERENCES

- [1] Maghilnan S, Rajesh Kumar M- "Sentiment Analysis on Speaker Specific Speech Data", 2017 International Conference on Intelligent Computing and Control (I2C2).
- [2] Kevin Tomba, Joen Dumoulin, Elena Mugellini, Omar Abou Khaled and Salah Hawila- "Stress detection through speech analysis", 2018.
- [3] B. Padmaja, V.V.Rama Prasad and K.V.N. Sunitha "A Machine Learning Approach for Stress Detection using a Wireless Physical Activity Tracker", 2018.
- [4] Anderson R. Avila, Shruti R. Kshirsagar, Abhishek Tiwari, Daniel Lafond, Douglas O'Shaughnessy and Tiago H. Falk- "Speech-Based Stress Classification Based On Modulation Spectral Features And Convolutional Neural Networks", 2019 27th European Signal Processing Conference (EUSIPCO).
- [5] Ravinder Ahuja, Alisha Banga – "Mental Stress Detection in University Students using Machine learning", 2019.
- [6] Zhao Cheng Huang, Julien Epps, Dale Joachim, Vidhya Saharan Sethu, IEEE Journal of Selected Topics in – "Signal Processing, Natural Language Processing Methods in Speech-based Stress Detection for Acoustic and Landmark Event-based Features", 2019.
- [7] Russell Li, Zhidong Liu- "Stress Detection Using Deep Neural Networks", 2020.
- [8] Dr. S. Vaikole, S. Mulajkar, A. More, P. Jayaswal, S. Dhas- "Stress Detection through Speech Analysis using Machine Learning", Volume 8, Issue 5 May 2020.
- [9] Zarinatj Hosseinzadeh-Shanjanji, Khadijeh Haji Miri, Bahram Rostami, Shokoufeh Ramezani, Mohsen Dadshi, -"Stress Anxiety Levels Among