International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 06 | June 2020 www.irjet.net p-ISSN: 2395-0072

Implementation of Mood Detection through Voice Analysis using Librosa and CNN

Mr. Hrushikesh Mohanty[1], Ms. Shivani Budhvant[2], Mr. Parag Gawde[3], Prof. Manjusha Shelke[4]

[1]Student, Department of Information Technology, Excelsior Education Society's K.C. College of Engineering & Management Studies & Research, Thane, Maharashtra, India
[2]Student & Department of Information Technology, Excelsior Education Society's K.C. College of Engineering & Management Studies & Research, Thane, Maharashtra, India
[3]Student & Department of Information Technology, Excelsior Education Society's K.C. College of Engineering & Management Studies & Research, Thane, Maharashtra, India
[4]Assistant Professor, Department of Information Technology, Excelsior Education Society's K.C. College of Engineering & Management Studies & Research, Thane, Maharashtra, India

---

Abstract - *Mental Health is one of the rising issues in today's stressful life. This issue has taken such serious turns that having a mental stability for an average person has become a luxury of sorts. Having an emotional, psychological and a social well-being is utterly important. An AI assistant that can understand the mood of the user and personalize things in such a way that the user feels some sort of peace of mind could be a factor that can contribute to mental well-being. Things like automatically changing the colours of the lights in the room or setting a good music to play in the background could create an ambience that might help to reduce the stress a person is going through. Such tasks could be performed by the AI when it interacts with user on a daily basis. For this, we proposed an integration of a module for a voice assistant that interacts with the users and analyses their sentiment through that recorded voice. The results of the analysis is then used to perform tasks like automatically changing the colour settings of the lights in the room and play a suitable music in the background. We have explored through various domains such as speech recognition, voice detection and machine learning models to extract the exact idea and derive utmost clarity in our project implementation. We used a CNN model to train on sample data of around 2000 audio files which was then used to record voice and give an output stating the current mood of the speaker.*

Key Words: **Artificial Intelligence (AI), Convolutional Neural Networks (CNN), Mel Frequency Cepstral Coefficient (MFCC), Machine Learning (ML), Sentiment Analysis, Voice Recognition.**

1. INTRODUCTION

There have been huge advancements in research regarding Machine Learning and Neural Network fields. In the recent few years, with a massive development in the field of Internet and social-networking, people have started giving out their opinions and expressing their emotions online more freely than ever. Sentiment Analysis explores this behaviour of humans and tries to figure out the general emotion expressed by a massive amount of people. Sentiment Analysis uses the concept of Natural Language Processing where it performs actions like analyse, reason and give the text an emotional score or colour [1]. Although sentiment analysis has been performed on texts plenty of times, it is the idea of performing it on voices that really drove us into this project.

A person's emotions and mood swings are directly correlated to their surroundings and circumstances. Listening to good music, having bright and better lighting conditions have been proven to change the mood of a person for better and helps in avoiding negative thoughts that lead to stress. With the rising concern for mental health in modern times due to an average person's stressful schedule, it has become an important task to keep a tab on thoughts and emotions felt by a person. Several AI systems have started working on the field of automation based on individual behaviours and work patterns. Detecting the mood and emotions of a user with their voice opens up doors to limitless Artificial Intelligence automation technology and personalization possibilities. With the already existing AI systems based on voice detection and speech analysis, the integration of emotion detection increases the functionality of the AI by many folds. Our proposal was based on this very idea of managing certain manual tasks automatically by judging the mental state of the user.

Plenty of libraries are available to record and store an audio file to perform analysis and different functions on it but the most popular one is said to be Librosa. Using speech detection and sentiment analysis properties along with correct machine learning models trained on a large amount of data set increases the accuracy with which the system detects the correct mood of the speaker. CNN has the ability to combine feature extraction and classification tasks making it a more accurate than other models out there [2].

2. BACKGROUND AND RELATED WORK

Studies in the field of sentiment analysis have been done plenty of times on text based forms of human speech in

various fields of implementation. A typical project was implemented in 2016 where sentiment analysis was used to analyse the sentiment of Twitter users by incorporating sentiment-specific word embedding (SSWE) and weighted text feature model (WTFM) [3]. Another study in 2018 showed that analysing texts using CNNs combined with SVM text sentiment analysis yielded better accuracy results of sentiment classification when compared to traditional CNN models [4].

Our model has taken references from Voice Recognition using MFCC Algorithm model (Koustav Chakraborty, Asmita Talele and Prof. Savitha Upadhya, 2014) [5] where they proposed a model which simulated the calculation of MFCC (Mel Frequency Cepstral Coefficient) from audio files. The simulation was done in a software called MATLAB R2013a [5]. This study showed that MFCC is a special feature in human speeches having their unique frequencies and can be used for speech recognition systems. The comparison resulted in conclusion that every user had a unique and different MFCC for their voices and varied with different moods. In our project, we proposed to use the MFCC feature while analysing the audio to predict the gender and mood of the speaker.

Another project we referred was Human Vocal Sentiment Analysis [6] where they used deep neural networks like CNNs with vocal feature extractions like the MFCC used in our project to predict mood of the speaker. This project was our key reference in creating a module for existing AI systems, like Amazon Alexa or Google Assistant, to perform tasks based on speaker's sentiment values and use them to personalize tasks accordingly. In this project, they have also introduced recognizing vocal features of a speaker using MFCC and STFT (Short Time Fourier Transform) using the RAVDESS dataset consisting of 8 emotion classes. We decided to include this dataset as the core dataset in our project.

3. METHODOLOGY

According to the official documentation, Librosa is a python package for music and video analysis. It provides the building blocks necessary to create music information retrieval systems. It also provides functionality properties like tuning and checking the sampling rate of the audio files it stores and records. Using Librosa, features from an audio file could be extracted like the MFCC (Mel frequency coefficient) which is similar to 'edges' that are a feature extracted from images [7]. MFCC considers human perception for sensitivity at appropriate frequencies by converting the conventional frequency to Mel Scale. Thus, for this reason itself, MFCCs are suitable for speech recognition tasks quite well. It represents distinct units of sound or phenomes, as they are known, as the shape of vocal tract, responsible for generation of sound, and are suitable to understand humans and frequencies at which humans utter or speak sound [7]. These would became a major feature that the ML model we built detected and used to classify the audio files accordingly.

To be able to extract correct features from an audio file and relate it to the output emotion, a large amount of dataset should be used as training data to train a ML model. Our requirement of the dataset was that it should contain correct amount of recognizable emotions expressed by humans through their voices with distinct and clear sound. The RAVDESS dataset [9] seemed to be the correct choice as it contained speech data which was available in three different formats, viz. audio-visual, audio-only and visual-only. Second type of file, audio-only would be the correct choice as our project dealt with finding emotions from speech and had nothing to do with visuals. 1500 audio files were used from RAVDESS which were from 24 different actors and in 'wav' format. The second set of audio files [10] had around 500 audio speeches from four different actors with different emotions. Each of the audio file had a unique identifier at the 6th position of the file name which was used to determine the emotion of the audio files inside the folder. For example, a particular audio from 'Actor 01' folder of the dataset had the file name '03-01-01-01-01-01-01.wav'. Here the 01 in the 6th position stood for the first emotion 'calm'. Whereas the 6th identifier of the file '03-01-02-02-02-02-04.wav' from 'Actor 04' folder identified the second emotion 'happy'. The goal of the project was to determine an emotion out of a range of 5 different emotions; these were 'calm', 'happy', 'sad', 'angry' and 'fearful'. The dataset was then imported and loaded to be used in our project. A list variable was created to store the key-value pairs of emotions along with their correct identifiers. Separating out the voices of male and female speeches resulted in an increased accuracy rate which was found later on in the project. This would then require us to create a new list which had double the key-value pairs to identify the emotion of male and female separately. We tuned the dataset to our needs like increasing the sampling rate, shuffling the data and splitting it into train and test set for the model to be trained. We also tested out the working of Librosa library by plotting the spectrograms and waveforms of the audio files to try to know its features. Each of the audio files gave us plenty of features which were nothing but array of many values. These features were then appended by the labels which we had created in the previous steps.

We noticed that some audio files had missing features as they were shorter in length. This resulted in our next step being to deal with these missing features. For this, we decided to increase the sampling rate by twice to get the unique features of each emotional speech. Increasing the sampling frequency even more would have resulted in collection of noise and thus distorting the audio file which would have affected the results of the prediction in a negative manner.

We initially proposed to work and build a LSTM model to train our dataset. LSTM is a Recurrent Neural Network which can learn both long term and short term dependencies

in data [11]. This neural network model consists of multiple cells and each has three main components which perform the function of forgetting, remembering and updating data [11]. But the LSTM model showed a low accuracy rate which went against our goal. After this we proposed to build a CNN model. A CNN uses perceptron, a machine learning unit algorithm used for supervised learning, to analyze data. CNNs are widely used in classification models built for images, natural language processing and other kinds of cognitive tasks. In our project we used 1D CNNS as these are counterparts to the traditional CNN and work on 1D signals, like in audio files, which have limited labeled data and high signal variations [2]. As our project is a classification problem where the audio files must be classified into different emotions, CNN worked best for us by giving us a higher accuracy score with a satisfactory number of tunings and modifications. The features of the audio files were one dimensional array which is not suitable for the traditional CNN libraries to work on. We used available Python libraries of Conv1D, MaxPooling1D and Average1D for all the layers of the CNN. The numbers of CNN layers were tuned to suit the efficiency of the model and the machine the model would be trained on. After the model was trained, the test data which we had split earlier was tested to validate the accuracy and prediction and we checked if the model didn't have any errors or anomalies, particularly, 'overfitting'.

Our model predicts the voice recorded by the speaker which is stored and analysed and then tested through the trained ML model. It has been observed that different colours have powerful impact on emotions of a human being. Certain colours induce certain kind of behaviours in humans, and this is the reason why restaurants or cafes are lit up in a certain way [12]. As we have consider to work on 5 emotions as of now, our home automation system changes the light setting of the rooms based on the same. Following shows the changes in light settings according to the predicted emotions [13]:

- If the mood detected is anger the light setting will change to blue colour as it brings peace to mind.
- If the mood predicted in user's voice is sad or fearful, the light settings changes to colour yellow.
- If the emotion is predicted to be calm, light settings would change to purple colour as it is said to reduce mental stress and puts you at ease.
- If the emotion is predicted to be happy, green would be the colour of choice as it increases creativity in human mind.

Figure-1: Blue light helps in peace of mind (Begoodesign.net)

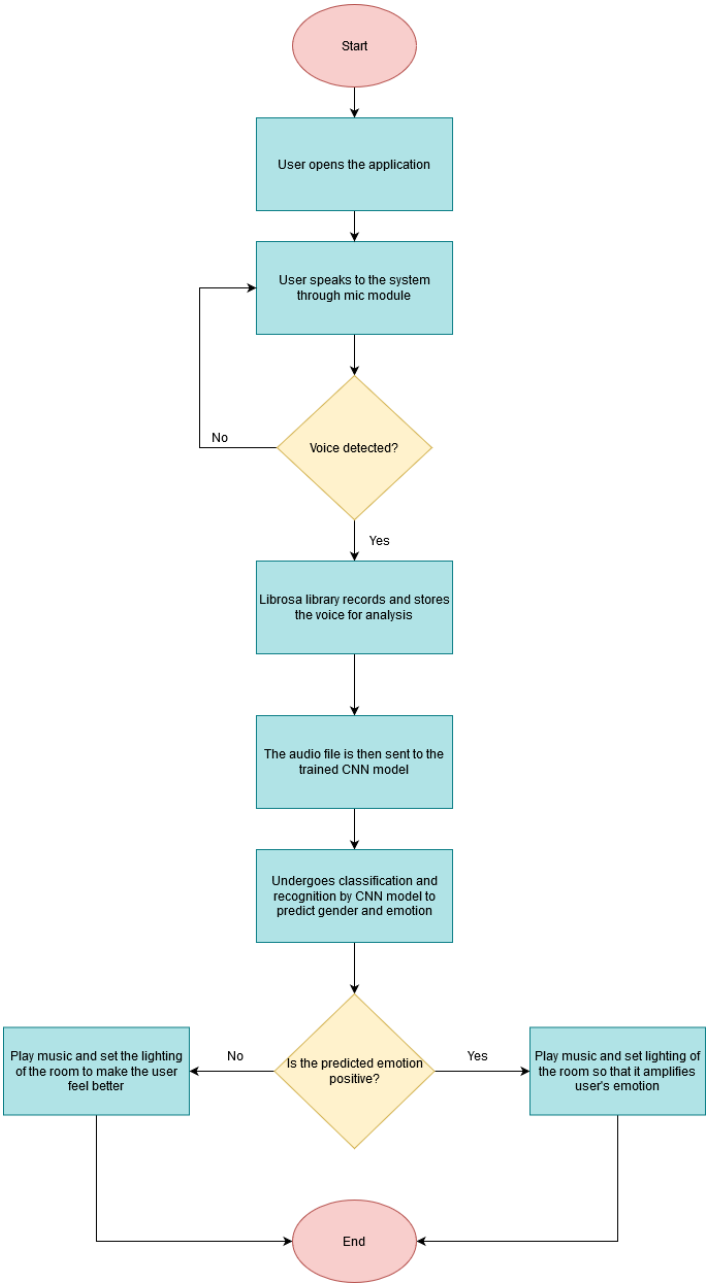

Figure-2: Basic flow of our model

We also built a dataset of music from different genre according to the five moods. When the model detects the user's mood the model fetches music from the created dataset for the particular mood. If the emotion detected is a negative emotion like sad, anger or fearful, the model would play songs to change such emotions. If the emotion detected is a positive emotion like happy or calm, the model would play songs from the dataset to amplify such emotions.

4. ANALYSIS AND RESULT

At first, we decided to build a prediction model using LSTM neural network. The LSTM neural net gave a low training accuracy of about 15% when we used 5 layers with batch size of 32 and 50 epochs. Due to the low accuracy rate we decided to look for better prospects to build our neural net.

Thus, we moved onto CNN model, which we tuned for best results after experimenting with the model. For best results, we found that a model with 18 layers, 'softmax' function, 'rmsproop' activation function, a batch size of 32 and 1000 epochs gave optimum results. Also, segregating the dataset by male and female voices increased the accuracy rate by 15%. The final validation accuracy rate achieved by the model was 60%.

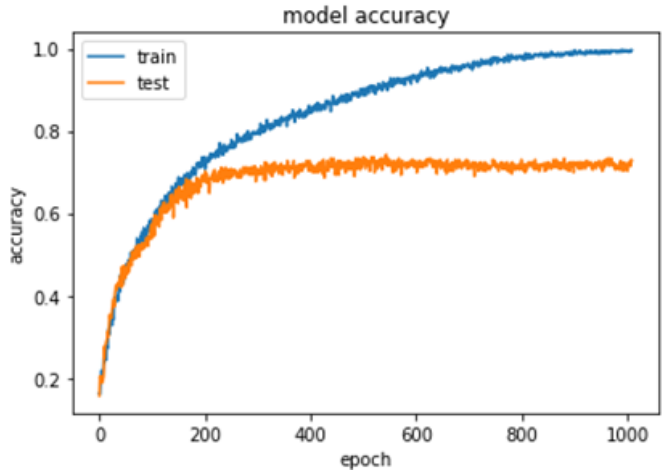

Figure-3: Accuracy of CNN model

With the above modifications to the neural net, we started getting efficient results and the model started giving fairly correct predictions.

5. CONCLUSION

The ML model which we built was based on a relatively new idea where there has been comparatively lesser research. This project can be extended further by adding more modules to make it more efficient and add new AI functionalities to enhance the overall experience of the user. At the moment, our model has 60% accuracy, which could be increased by adding more varied and vivid dataset and using better functions in the layers of the neural network.

In future, we plan to build modules where the AI tracks the emotions of the user to pre-plan an entire day's schedule including the pass-times which will be subjected the user's behaviour pattern.

REFERENCES

[1] Chen, Y., & Zhang, Z. (2018). Research on text sentiment analysis based on CNNs and SVM. 2018 13th IEEE Conference on Industrial Electronics and Applications (ICIEA).

[2] S. Kiranyaz, T. Ince, O. Abdeljaber, O. Avci and M. Gabbouj, "1-D Convolutional Neural Networks for Signal Processing Applications," *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, United Kingdom, 2019, pp. 8360-8364.

[3] Li, Q., Shah, S., Fang, R., Nourbakhsh, A., & Liu, X. (2016). *Tweet Sentiment Analysis by Incorporating Sentiment-Specific Word Embedding and Weighted Text Features. 2016 IEEE/WIC/ACM International Conference on Web Intelligence (WI).*

[4] Chen, Y., & Zhang, Z. (2018). *Research on text sentiment analysis based on CNNs and SVM. 2018 13th IEEE Conference on Industrial Electronics and Applications (ICIEA).*

[5] Koustav Chakraborty, Asmita Talele, Prof. Savitha Upadhya (2014), *Voice Recognition using MFCC Algorithm. International Journal of Innovative Research in Advanced Engineering (IJIRAE), Volume 1 Issue 10 (November 2014).*

[6] Andrew Huang , Martin (Puwei) Bao, *Human Vocal Sentiment Analysis* (2019), [Online] Available: https://arxiv.org/abs/1905.08632

[7] https://medium.com/prathena/the-dummys-guide-to-mfcc-aceab2450fd

[8] Hajiaghayi, M., & Vahedi, E. (2019). Code Failure Prediction and Pattern Extraction Using LSTM Networks. 2019 IEEE Fifth International Conference on Big Data Computing Service and Applications (BigDataService).

[9] http://neuron.arts.ryerson.ca/ravdess/?f=3

[10] http://kahlan.eps.surrey.ac.uk/savee/Download.html

[11] https://medium.com/@RaghavPrabhu/understanding-of-convolutional-neural-network-cnn-deep-learning-99760835f148

[12] Sroykham, W., Wongsathikun, J., & Wongsawat, Y. (2014). The effects of perceiving color in living environment on QEEG, Oxygen saturation, pulse rate, and emotion regulation in humans. 2014 36th.

[13] https://www.dmlights.com/blog/effect-coloured-light-on-human-body/