International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-9, July 2020

Speech Based Depression Detection using Convolution Neural Networks

Swathy Krishna, Anju J

Abstract*: Depression has become a serious mental disorder nowadays affecting people of almost all age groups. Loss of interest in daily activities, constant feeling of being isolated, hopelessness etc. causing significant impairment in life. This illness will affect the physical and mental health of the individual affecting his/her emotional stability. Emotions are a way of expression of one's state of mind in the form of thoughts, feelings or behavioural responses. For a depressed individual, the emotions are often negative in nature. Diagnosis of depression is a complex task as the disease may be unidentified by the patient itself. Sometimes the patient may be reluctant to consult a doctor. The long term ignorance of the illness may worsen the mental health of the one suffering from it. Thus the early diagnosis of depression is of great significance. With the emergence of neural networks and pattern recognition, many researchers have put effort in detecting depression by analysing non-verbal cues, such as facial expressions, gesture, body language and tone of voice. Recent studies have shown that the speech emotion analysis can effectively be used in distinguishing emotional features and a depressed speech varies from that of a normal speech to a great extent. The depressed patient normally speaks in a low voice, slowly, sometimes stuttering, whispering, trying several times before they speak up or become mute in the middle of a sentence. This paper proposes a CNN architecture for learning the audio features in the speech for detecting depression, identifying the emotions and to infer the emotional severity of the individual. This paper also reviews some of the existing research methods in the field of depression analysis.

Keywords: CNN, Depression, Spectrogram, Speech

I. INTRODUCTION

Depression is a serious mood disorder which is often undiagnosed in the early stages. If left untreated it may even lead to severe health conditions, suicide etc. Many people are affected by depression making them less sociable, thus making the diagnosis difficult. Some of the common depression scales that is used for diagnosis are PHQ-8[1], BDI-2, GDS, etc. The traditional approach for the diagnosis of depression is based on conducting clinical interviews to screen candidates for depression. But, these evaluations are highly depend on the questions administered by the clinician. There is a necessity to develop automatic techniques for supporting the diagnosis of depression. Recently with the emergence of machine learning and artificial neural networks, several methods have been developed in recent years for supporting clinicians during the diagnosis and monitoring of clinical depression.

The evaluation of patients with depression depends on two factors—clinical history (i.e., history of presenting symptoms, prior episodes, family history etc.) and the mental state examination (appearance, speech, movement). The latter should be focused more for the effective diagnosis. In particular, the analysis of audio data is found to be much effective for assessing the mental state.

Behaviors, poses, actions, speech and facial expressions; these are considered as channels that convey human emotions. Emotions are an important part of human life and many researches has been carried out to explore the speech emotions [2]. Fluctuations in mood are a very common thing that can happen to anyone. These mood changes may sometimes last for hours, days or even weeks. If it persists for a long term and affects the person's personal and social lives, the chances of the person having depression are very high. People affected from depression often feel isolated, have decreased social contact, prolonged feel of sadness, loss of interest in daily activities etc. Speech is an emotion which truly identifies one's mental state. An important issue in the process of a speech emotion recognition system is the extraction of suitable features that appropriately characterize the variation in the emotions. Since pattern recognition techniques are rarely independent of the problem domain, it is believed that a proper selection of features significantly affects the classification performance. The emergence of neural networks have significantly increased the classification accuracy in depression detection. The neural networks learn from the patterns of the input features that can clearly distinguish the spectral characteristics of normal and depressed groups thus improving classification accuracy. This paper proposes the use of deep neural networks in detecting the depression using speech features using DAIC-WOZ [3]. It discusses the importance of spectral features in identifying the voice features. It also identifies the emotional severity of the individual. The remaining part of the paper is organized as follows : Section II reviews some of the existing approaches in the field of depression detection and section III briefs the overview of proposed system. Section IV highlights the process of detecting depression using speech and section V how the emotions are inferred from the speech samples.

Revised Manuscript Received on July 30, 2020.
***** Correspondence Author
Swathy Krishna*, Department of Computer Science, Lal Bahadur Shastri Institute of Technology for Women (LBSITW), Thiruvananthapuram, Kerala, India.
Anju J, Assistant Professor, Department of Computer Science, Lal Bahadur Shastri Institute of Technology for Women (LBSITW), Thiruvananthapuram, Kerala, India.

Finally the section VI discusses the procedures carried out and results.

II. RELATED WORKS

Depression is considered as a common mental disorder which affects the mental health of the person making it complex to diagnose. Emotional outbursts are usual in depressed individuals. They often deal with emotional instabilities. So identifying the variations in emotions is very crucial in depression assessment. Recognising the emotions helps in exploring the different factors of depression. Emotions are expressions that reflect the mental state of the person. Studies have shown that the speech contain many peculiar features for detecting the mental state of a person. Quatieri and Malyska [4] came up with the findings that voice of an individual contains information about the mental state. They suggests that the voice features can be used for analysing depression severity. Features like jitter, shimmer, amplitude variation in voice regions can be effectively utilized in detecting depression. Speech analysis is also widely used in screening participants during interviews for measuring stress. Kevin Tomba, Joel Dumoulin, Elena Mugellini , Omar Abou Khaled and Salah Hawila [5] uses the mean energy, the mean intensity and Mel- Frequency Cepstral Coefficients (MFCCs) as features to determine if the interviewed candidates are stressed or not. The Berlin Emotional Database (EmoDB), the Keio University Japanese Emotional Speech Database (KeioESD) and the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) are used as datasets. It compares the results of both Support Vector Machine and Artificial Neural Network for classification of emotions. The best results were obtained with neural networks. He Lang, and Cui Cao [6] uses a combination of handcrafted and deep learned features which can effectively measure the severity of depression from speech was used. Deep Convolutional Neural Networks (DCNN) are built to learn the deep learned feature from spectrograms and raw speech waveforms. Then median robust extended local binary patterns were extracted from spectrograms. The performance of depression recognition were improved by the introduction of adapt joint tuning layers, to combine the raw and spectrogram DCNN.

Alghowinem et al. [7] evaluates discriminative power of read versus spontaneous speech (in an interview / conversation) in detecting depression. Support Vector Machines (SVM) was used as classifier. The results show that the spontaneous rate has more variability which increases the recognition rate of depression and the MFCC play a significant role in the classification. The emotions are conveyed through different channels such as face, speech, body behaviours, head movements etc. Shubham Dham, Anirudh Sharma and Abhinav Dhall [8] utilizes text, audio and visual data in DAIC- WOZ database for depression detection. Gaussian Mixture Model (GMM) clustering and Fisher vector approach were applied on the visual data, low level audio features and head pose and text features were also extracted. SVM classifier was applied separately on the extracted features. Finally decision level fusion was used for combining the results of different modalities. Karol Chlasta, Krzysztof Wołk and Izabela Krejtz [9] proposes a novel method to detect depression using deep convolutional neural networks. The experiments were done on Distress Analysis Interview Corpus (DAIC) and uses ResNet-34 for the classification. The results suggests that audio spectrograms are a promising feature for screening of depressive subjects. And also the use of short voice samples reduced the effect of noise. So the analysis of emotions of an individual is of great concern in the case of depressed individuals. The universal emotions are categorised into seven such as happy, sad, disgust, contempt, neutral, angry and fear. The studies have shown that the depressed individuals frequently have a negativity in their emotions. In the past few decades, emotion recognition has been an active area of research which has applications in the field of facial tracking, automating video analytics, robotics etc. As depressed individuals often find it difficult to communicate their symptoms to others, the analysis of speech helps in identifying one's true emotions. The depressed speech has characteristics such as reduced speaking intensity, reduced pitch range, slower speech, etc. Since depression is a complex disorder, analysing it from a particular aspect may not be sufficient for the effective assessment. The emergence of deep neural networks has gained attention in the field of speech analysis as it can learn the features of the input by itself.

III. SYSTEM OVERVIEW

The project presents a framework for detecting depression using speech from spectral features and then analysing the emotions in the speech. The speech (wav) samples are trained using a Convolutional Neural Network model.

Initially the troubled recordings in the dataset should be identified and removed as these are unsuitable for further feature extraction.

Then segmentation is carried out for separating the participant's speech from other speakers, noise, silence etc. The spectrogram is generated for the audio samples and given as input to the convolutional neural network for feature extraction.

After the classification, the emotions in the speech should be analysed. If the person is classified as depressed, the emotions in speech such as fear, anger, disgust etc. which are called as the negative emotions should be identified. As the depressed individuals are emotional, the analysis of negativity of the emotions is of great importance.

If the person is classified non depressed, there is no point in analysing the negative emotions, and the positive emotions in the speech should be identified. So the proposed system has two major parts, depression detection and emotion analysis.

IV. DEPRESSION DETECTION

A. Database Description

DAIC- WOZ database contains clinical interviews that are intended to support the diagnosis of mental health conditions like depression.

The interviews were conducted by a virtual agent called Ellie as shown in Fig. 1, which were controlled by a human interviewer in another room.

Fig. 1. Ellie, the virtual interviewer

The database were published by University of California for supporting the research purposes. The archive consists of 189 folders of participant's interview sessions. Each folder corresponds to the information on the interview session of participant i.e. audio recordings, video recordings, textual transcriptions etc. It consists of 30 depressed subjects and 77 non-depressed subjects. All audio files (wav format) were recorded at 16 kHz. The duration of session varies from 7 – 33 minutes (average of 16 mnts). The class labels was assigned on the basis of PHQ –8 score. If the PHQ-8 score is greater than 10, then the participant is depressed else he or she is non-depressed.

B. Data Preprocessing

As the interviews are recorded in the real time environment, there are chances of technical issues while recording the interviews. Some recordings are not suitable for further feature extraction due to excessive static, proximity to the virtual interviewer, volume levels, etc. The first step in analyzing a person's acoustic features of speech is segmenting the person's speech from silence, other speakers, and noise. Fortunately, the participants in the DAIC-WOZ study were wearing close proximity microphones in low noise environments, which allowed for fairly complete segmentation using the pyAudioanalysis module. After segmentation, the segmented wav files are converted to a single wav file with the vast majority of silence and virtual interviewer speech removed. Feature extraction is later performed on these wav files. In DAIC WOZ, the number of non-depressed subjects is four times than the number of depressed subjects. So to address the class imbalance issue and to increase the size of training data, audio slicing was introduced. The preprocessed audio file was sliced into equal slices. This also normalize the duration of the training samples.

C. Spectrogram Generation

A spectrogram also called sonographs is a visual representation of spectrum of frequencies of a signal as it varies with time. The spectral features (frequency based features), which are obtained by converting the time based signal into the frequency domain using the Fourier Transform are used for analysing the speech features. Spectral features play a significant role in analyzing the energy variations, amplitude etc. Spectrogram is generated for all the audio files in the enlarged dataset. An example of the generated spectrogram is shown in Fig. 2.

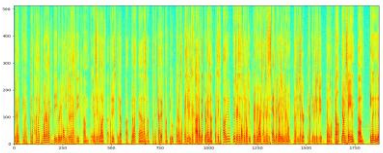

Fig. 2. Spectrogram of an audio

D. Feature Extraction Using CNN

Feature extraction is carried out with the help of deep neural network architecture. CNNs take images as input. Thus the audio samples are converted to spectrogram images. A filter (kernel) is subsequently slid over the spectrogram image and patterns for depressed and non-depressed individuals are learned. The CNN begins by learning features like vertical lines, but in subsequent layers, begins to pick up on features like the shape of frequency-time curve (perhaps representative of speaker intonation). Such learned features may provide an elegant and powerful representation of speech features, which in turn are representative of underlying differences between depressed and non-depressed speech.

V. INFERING EMOTIONAL SEVERITY

After the depression detection using speech, the severity of emotions in the speech should be identified. Because a depressed individual will mostly have emotional instability. They mostly look sad, speak in low voice, short tempered etc. The universal emotions are happy, surprise, disgust, sadness, anger and fear. Studies have shown that depressed individuals often have negative emotions. So models have been developed for both depressed and non-depressed for identifying positive and negative emotions. If the person is classified as depressed, the emotions in speech such as fear, anger, disgust etc. which are called as the negative emotions should be identified. As the depressed individuals are emotional, the analysis of negativity of the emotions is of great importance. If the person is classified non-depressed, there is no point in analysing the negative emotions, and the positive emotions in the speech should be identified. Finally a probability graph is generated for analysing the severity of emotions. For emotion recognition SAVEE (Surrey Audio-Visual Expressed Emotion) [10] is used. It is an emotion recognition dataset consisting of recordings from 4 male actors in 7 different emotions, 480 British English utterances in total. It contain the audio files expressing happy, sad, surprise, disgust, fear and anger. Spectrograms are generated for each file in the emotion classes. These are used for further feature extraction process. The spectrograms of the audio files of different audio class are given as input to the CNN for further classification. The depressed and non-depressed models are created for predicting the emotional severity of depressed and non-depressed individuals. Studies show that a depressed individual have a frequent outburst of negative emotions. So the severity of negative emotions such as sad, anger, disgust, fear is analysed by the depressed model. The non-depressed model is used to analyse the positive emotions such as happy, surprise etc.

VI. RESULTS AND DISCUSSIONS

The experiments were done on Google Colaboratory platform which enabled us to use NVidia K100 graphic cards for computations. The depression detection using speech was done on DAIC- WOZ database. The spectral features of the audio is exploited for identifying the patterns that distinguish depressed and normal groups. Since the dataset suffers from class imbalance problem, the experiments were done on the extended dataset which was generated by slicing the audio files into equal chunks with a duration of 1 minute. The network is trained with the spectrogram images of both depressed and non-depressed class respectively. The last layer of the CNN, i.e. softmax layer is used for the classification purpose. The input images are resized to 224 * 224. The batch size is 128. The number of epochs is 5. Testing is carried out by inputting new audio files and analysing the class labels predicted by the model. The proposed model has been successful in detecting the depressed and non-depressed using speech features. The proposed system also analyses the emotions of the depressed and non-depressed.

VII. CONCLUSION

This project addresses the depression detection by exploiting the various spectral features in the speech. In the past few years, many researchers have come up with different approaches for detecting depression. We use CNN model for learning the high level features in the speech by extracting features from the spectrogram, a visual representation of the spectrum of frequencies of an audio signal. It has been found that deep learning approaches are quite effective in learning the patterns of the depressed and non-depressed individuals from the spectrograms. As depressed individuals are often sensitive and emotional, the study of negativity in their emotions is of great importance. The system is also capable of analysing the emotion severity in the predicted classes. In the long term, the system can be used for supporting clinicians in depression analysis.

REFERENCES

1. Kroenke, Kurt, et al. "The PHQ-8 as a measure of current depression in the general population." *Journal of affective disorders* 114.1-3 (2009): 163-173.
2. Sailunaz, Kashfia, et al. "Emotion detection from text and speech: a survey." *Social Network Analysis and Mining* 8.1 (2018): 28.
3. Gratch, Jonathan, et al. "The distress analysis interview corpus of human and computer interviews." LREC. 2014.
4. Quatieri, Thomas F., and Nicolas Malyska. "Vocal-source biomarkers for depression: A link to psychomotor activity." *Thirteenth Annual Conference of the International Speech Communication Association*. 2012.
5. Tomba, Kevin, et al. "Stress Detection Through Speech Analysis." *ICETE (1)*. 2018.
6. He, Lang, and Cui Cao. "Automated depression analysis using convolutional neural networks from speech." *Journal of biomedical informatics* 83 (2018): 103-111.
7. Alghowinem, Sharifa, et al. "Detecting depression: a comparison between spontaneous and read speech." *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013.
8. Dham, Shubham, Anirudh Sharma, and Abhinav Dhall. "Depression scale recognition from audio, visual and text analysis." *arXiv preprint arXiv:1709.05865* (2017).
9. Chlasta, Karol, Krzysztof Wołk, and Izabela Krejtz. "Automated speech-based screening of depression using deep convolutional neural networks." *arXiv preprint arXiv:1912.01115* (2019).
10. Jackson, P., and S. Haq. "Surrey audio-visual expressed emotion (savee) database." *University of Surrey: Guildford, UK* (2014).

AUTHORS PROFILE

Swathy Krishna, is currently pursing M-Tech degree in Computer Science and Engineering at Lal Bahadur Shastri Institute of Technology for Women (LBSITW), Thiruvananthapuram, Kerala affiliated to APJ Abdul Kalam Technological University. She completed graduation in Computer Science and Engineering from Amal Jyothi College of Engineering, Kanjirappally, Kerala. Areas of interest are Image Processing, Machine Learning and Cyber security.

Anju J, is currently working as Assistant Professor in Computer Science department at at Lal Bahadur Shastri Institute of Technology for Women (LBSITW), Thiruvananthapuram, Kerala. Her areas of interest includes Mobile Communication, Sensor networks and Data Analysis. She has more than 5 years of teaching experience. She has published and presented many research papers in international journals. She has also attended many workshops on subjects such as Android Application, IoT and Cloud Architecture.