Depression Detection using speech as Input Signal

Aniket Waghela[1] ,Prinkle Singharia[2], Bhavya Haria[3],Bhakti Sonawane[4]

[1]Department Computer Engineering, Shah and Anchor Kutchhi Engineering College, Mumbai University.
[2]Department Computer Engineering, Shah and Anchor Kutchhi Engineering College, Mumbai University.
[3]Department Computer Engineering, Shah and Anchor Kutchhi Engineering College, Mumbai University.
[4] Prof. Department Computer Engineering, Shah and Anchor Kutchhi Engineering College, Mumbai University.

Abstract

In this paper we propose a method for computer based depression detection. It is focusing on two aspects: gathering conversations and getting only the patient's audio and creating a deep learning model for automatic depression detection. Convolutional Neural Network (CNN) classifier is used to find patterns in audio characteristics of the depressed patients. This training is carried out on DAIC-WOZ Dataset from USC's Institute of Creative Technologies and it was released as part of the AVEC(Audio/Visual Emotional Challenge) 2016. There is a sample imbalance in the dataset, as there are more number of non-depressed patients in the dataset. To remove this sample imbalance, we introduce random sampling before the model training.

Keywords: *Heart Disease, Data Mining, Decision Tree, K Nearest Neighbor, Support Vector Machine, Average K Nearest Neighbor, Genetic Algorithm, Artificial Neural Network.*

1. Introduction

Depression also known as Major Depressive Disorder is a serious and common medical illness that has negative effects on the way you think, how you feel and the way you respond. Fortunately, depression can be cured with treatment. Depression symptoms include persistent sadness and can be accompanied with declined interest in activities formerly enjoyed. Depression can lead to people losing their efficiency of functioning at home and at the workplace. Depression leads people to suicide and we lose 10.9 thousand for every 1 Lakh people each year in India [7].

This can be reduced by having a method for early detection, which can be done by designing a better system for depression detection. The system designed takes speech as input and checks for pattern similarity with the patterns found in patients already detected with depression. Learning of the system is carried out by implementing machine learning. Such systems will help greatly to speed up and early detection of this serious medical illness.

Depression detection in patients is normally done by visiting a doctor and carrying out a series of tests as prescribed by the doctor. These tests include- physical assessments used to rule out other medical causes of depression. The doctor tries to find out a major health concern(using blood tests) that may be the reasons for symptoms of clinical depression, examples include hypothyroidism and hyperthyroidism. Injuries in the Central Nervous System may also lead to depression. Doctors can usually identify depression in patients by asking particular questions and the standard procedure of conducting a physical exam. To rule out other diagnoses, the doctor may ask to take certain lab tests. Other tests may include- MRI of the brain or CT Scan, Electrogram, Electroencephalogram [8].

There are Depression Screening tests carried out to detect depression. All the above depression detection methods require time and are not very cost effective. Thus depression detection is not feasible for everyone. Thus, research for time and cost effective and instant depression detection have been carried out since 2009. This is where depression detection by combining machine learning concepts is used. Machine Learning concepts can use voice and speech as input to carry out various types of analysis, one of them being using voice and speech signals to classify between depressed and non-depressed patients [3].

Automatic Depression Detection is a topic rarely discovered until 2009. A lot of research has been done to automate the depression detection process using audio or video or both of them. Earlier work on automatic depression detection used models like SVM, Naïve Bayes, Random Forest, etc [3] .

The dataset in different studies contained speech of different types like scientific, emotional, picture description, interview, reading [1]. Studies have been carried out using prosodic features in speech. Recent research in this topic made use of Deep Learning Models [9].

The proposed work focuses on depression detection using speech as input and is independent of the words said in any language since the prosodic features of speech are considered for depression detection. The training is done using a CNN model. Remaining paper is organized as , Section (2) describes in detail about speech processing and its applications before, Section (3) presents various methodology of speech processing, Section (4) presents a brief overview of studies done on this topic, Section (5) describes in detail about the structure of DAIC-WOZ dataset, Section (6) describes implementation steps to produce the said results, Section (7) presents results for the model trained and gives an overview of the training phase of the model, Section (8) presents conclusion for the work proposed.

2. Importance of Speech Processing

Speech Processing consists of the study of speech signals and the various methods to process these signals. Speech processing is carried out by converting the signals into digital form. Thus, speech processing can be considered a form of digital signal processing with a condition that the signals are speech signals. Characteristics of speech processing include: acquisition, manipulation, storage, transfer and output of speech signal.

Speech processing can be considered as the convergence of two domains -

- Natural Language Processing
- Digital Signal Processing.

Speech processing technology has applications in digital speech coding, text-to-speech synthesis, spoken language dialog systems, automatic speech recognition. Speech controlled systems may prove to be easier to work with as it allows devices to be controlled in a contact-free, not requiring any monitoring, fast and instinctive way. Speech processing incorporated into devices may facilitate establishment of intelligent devices like smart phones that interact verbally with users. Real world example of a very successful speech processing technology is Siri. Siri was the outcome of Cognitive Assistant that Learns and Organizes (CALO) venture inside the Defense Advanced Research Projects Agency's (DARPA) Personalized Assistant that Learns (PAL) program, the biggest AI venture in United States history (SRI, 2015). Speech recognition technology has become an integral part of our everyday lives, yet as innovations are advancing, more intelligent systems will be developed by researchers. Given current patterns, in years to come, speech recognition technology will become a fast-growing (and world-changing) subset of signal processing.

3. Methodology for Speech Processing

Speech Processing can be carried out chiefly in time and frequency domains as seen Fig I.

In the time domain representation, the vertical axis is the amplitude or voltage of the signal and the horizontal axis is time. To process speech signals in time domain and evaluate the parameters for important change in signal, we must create consecutive window frames by dividing the signal. The windowed portion is the region of interest which is to be processed and the signal outside this window is zeroed out. Let W(n) be window size and S(n) be the signal then -

- If W(n) is short then the speech properties of interest will change little within the window.
- If W(n) is long enough, then it allows to calculate all the desired parameters of the speech but the longer windows average out the random noise present.
- If the size of the window is medium then analysis of S(n) is periodically repeated.

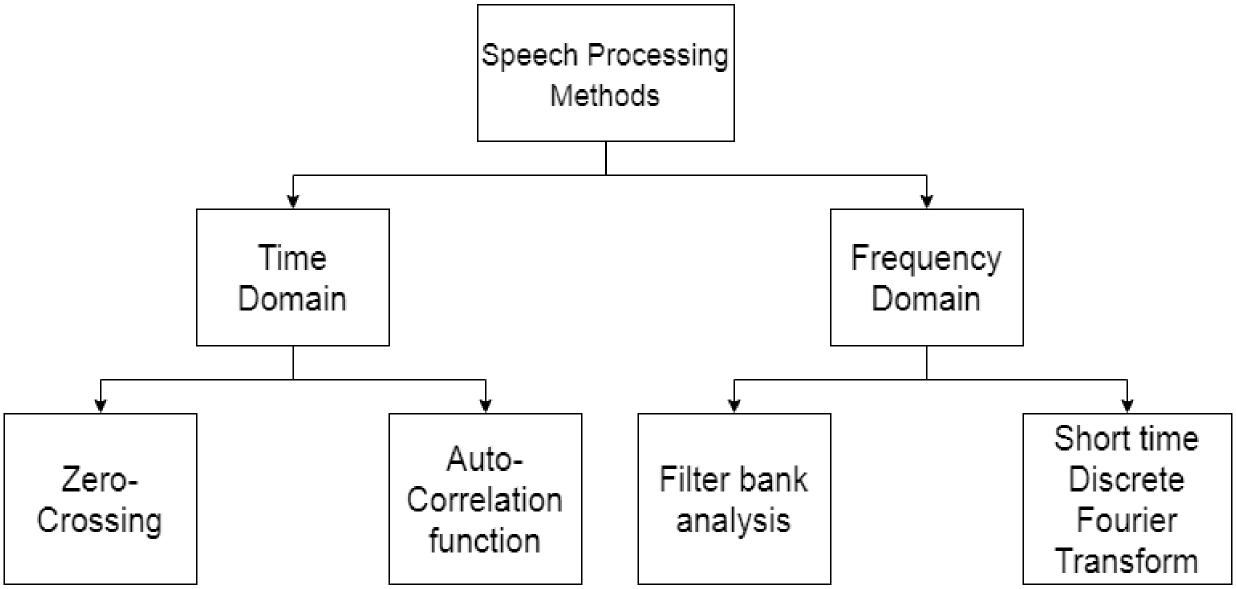
Speech Processing Methods
Time Domain
Frequency Domain
Zero-Crossing
Auto-Correlation function
Filter bank analysis
Short time Discrete Fourier Transform

Fig I. Methods of Speech Processing

The widely used time domain techniques are -

- Short-Time Averaging Zero-crossing Rate

Zero Crossing occurs in a signal when the algebraic sign of the signal changes or when the waveform crosses the time axis. For a sinusoidal signal, there are two zero-intersections per period.

- Short-Time Autocorrelation

The time signal r(k) for the inverse Fourier transform of the energy spectrum is called the autocorrelation function of S(n). The function r(k) preserves the information like the harmonic, periodicity and amplitude of S(n).

In the frequency domain representation, the vertical axis is still voltage but the horizontal axis is frequency. The different methods to extract the frequency domain parameters are -

- Filter bank Analysis

A filter bank is an array of band-pass filters that isolates the input signal into several number of segments, each one carrying a single frequency sub-band of the original signal (A band-pass filter is a device that allows frequencies to pass if they fall within a certain range and reject frequencies outside that range).

- Short Time Fourier Transform Analysis

The procedure for calculating the STFTs is -

1. Short and equal length segments of the longer time signal are created
2. Fourier transform of each of the smaller segments is calculated
3. The changing spectra as a function of time is then plotted(spectrograms).

4. Related Work

The study started when past discovery was done which suggested that our voice timbre contained information about our mood. Various studies were carried out that took into consideration the different features present in speech, the different types of speech, etc [11][12][13][14].

The speech processing can be divided mainly into linguistic and non-linguistic. The speech data that was collected in various researches was also of different kinds like emotional and scientific, positive, negative and neutral, natural conversation speech [1].

Extraction of prosodic features from speech vocal effects (formants), (pitch and energy) and glottal features facilitate the classification between depressed and healthy cases. Some research shows that voice characterization can be done by glottal features [15][16]. The velocity of airflow through the glottis from the lung is related to glottal features. In Fig II, glottal pulse sample is shown. For calculating the features of glottal pulse, use the algorithm shown in Fig III [3].

The dataset used by the authors in the research [3] was in Persian language. The dataset used included depressed and healthy treated students. This dataset was collected by the High school and expert psychologist. There were four groups in this database such that any of them includes 14 students in the age range of 16 to 18. First group contained depressed patients before the treatment. Second group consisted of the same patients but 2 months after they were given treatment. Third group consisted of depressed patients that did not receive any treatment. And the fourth group consisted of healthy cases . In this study, there were two types of text: Emotional text and Scientific text. This research used classification model Support Vector Machine (SVM) to classify if the patients are healthy or depressed cases.

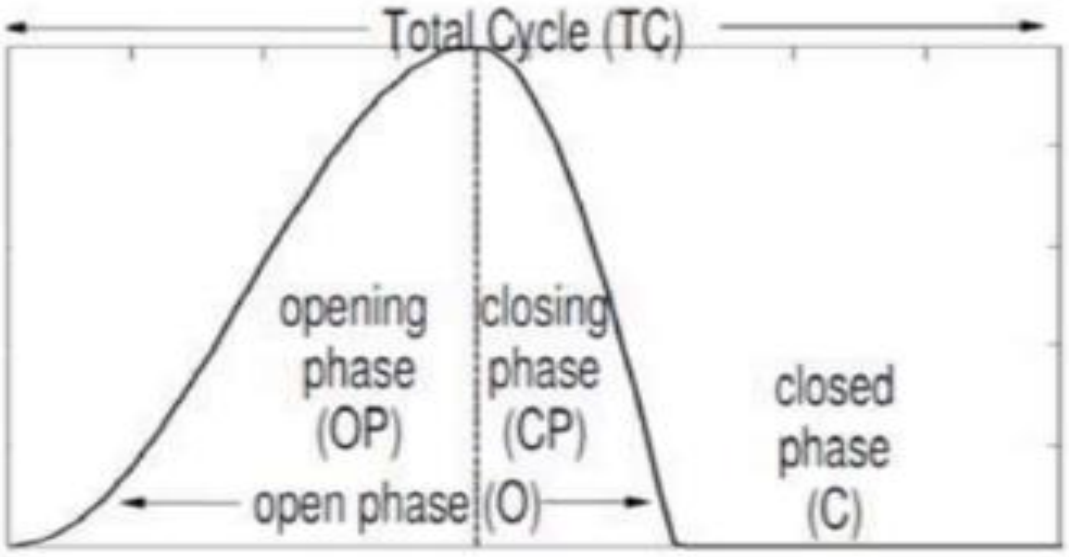

Fig II. The Glottal Pulse[3]

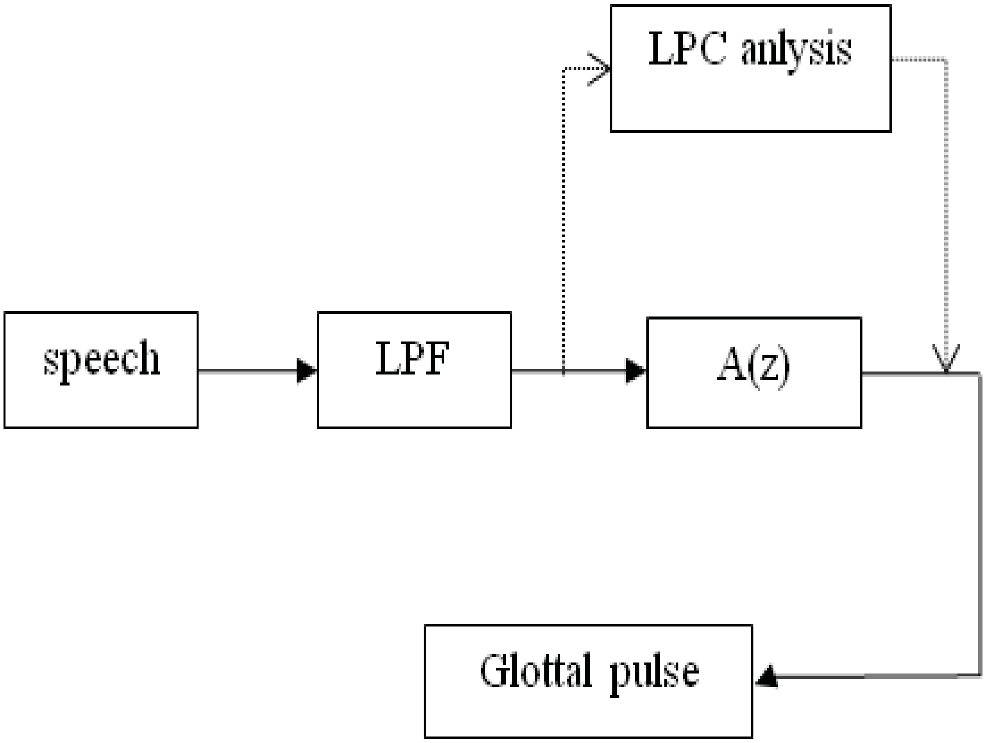

Fig III. The algorithm for extracting Glottal pulse [3]

Another approach used a multiple classifier system[1] as shown in Fig IV. The power to discriminate between different speech types was examined. Voice features like intensity, jitter,short time energy, loudness, zero crossing rate, shimmer, F0,MFCC, LPC,LSP and LPC were examined. The dataset used was created in Chinese language. The data collection included three parts - reading speech, interview speech and picture description speech, and each individual parts were further partitioned on the basis of emotions - positive, negative and neutral. Using all this data, for automatic depression detection, a new multiple classifier model was proposed. In the research it was found that speech from the interview gave better recognition rate than the speech from reading and picture description. In the experiment, only acoustic features were used so that depression detection can be done regardless of language. Three different classifiers were tested for their effectiveness. They are Support Vector Machine, Naive Bayes, Random Forest Classifier. After that, a new multiple classifier model that combined different speech types and emotions was used to classify between depressed and not depressed patients. This multiple classifier model is made of parallel topology consisting of many SVM's. This multiple classifier model has strength of each individual models, thus the performance was enhanced.

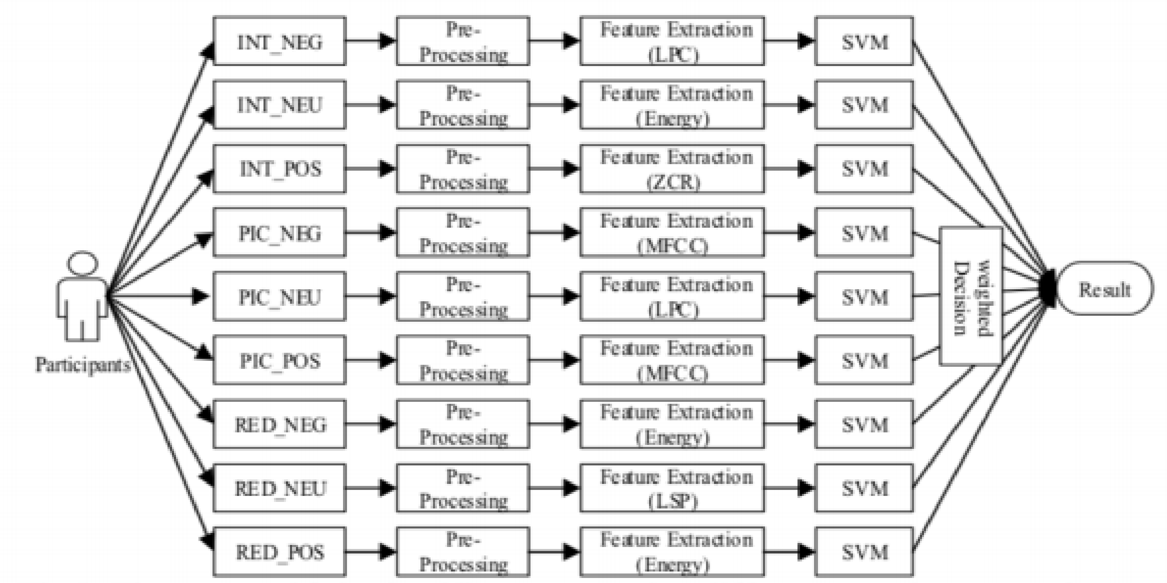

Fig IV. The new multiple classifier system using different speech types and emotions[1]

5. Speech Dataset

The dataset was provided by the DAIC(Distressed Analysis Interview Corpus)-WOZ Database created by USC's Institute of Creative Technologies and released as part of the 2016 Audio/Visual Emotional Challenge and Workshop. The dataset has 189 sessions containing an average of 16 minute sessions with a virtual interviewer named Ellie. For taking the interview , "Wizard of Oz" approach was used. In this the virtual interviewer was controlled by a human from the adjacent room. Each person who was interviewed was first checked for depression using a traditional approach. The dataset consists of Audio and Transcripts file of the interview taken for each interviewer. The main problem with the dataset was that the audio contained the voice of both interviewer and interviewee. Thus, pre-processing had to be performed to extract only the patient's speech. The dataset consists of audio and transcript files of the interviews.

6. Implementation Details

A. Pre-processing

The pre-processing aims at extraction of prosodic features to identify depression related characteristics. The major steps in preprocessing are Segmentation, and Noise removal. Segmentation includes removal of the interviewer's speech, and space gaps from the wav file for each interview to obtain separated wav files of an interview. Then the segmented audio files consisting of only the patient's voice are combined to obtain a single wav file for an interview. Segmentation has been carried out in an uncomplicated way using pyAudioAnalysis.

The noise in the audio signal was comparatively less since patients were wearing close proximity microphones, but this is not the case with speakers in mobile devices. Thus, further processing to remove noise has to be done for audio collected from mobile phones in real life scenarios.

B. Spectrogram Creation

A Spectrogram helps to visualize the spectrum of frequencies which are present in the input signal. Spectrograms are created by performing Short Time Fourier Transform of the signal. The digitally sampled signal is broken into chunks in the time domain, then Fourier Transformation is performed on these chunks individually to obtain the magnitude of frequency spectrum for each chunk. This transformation gives a measurement of magnitude versus frequency for a specific moment in time which is represented in the spectrogram by a vertical line. These chunks are then appended to form the final spectrogram image or a three-dimensional surface. The above process of creation of spectrogram corresponds to performing Short Time Fourier Transform(STFT) of the signal *s(t)*. That is, for a window width *ω,*

$$spectrogram(t, \omega) =|STFT(t, \omega)|\verb|^|2 \qquad (1)$$

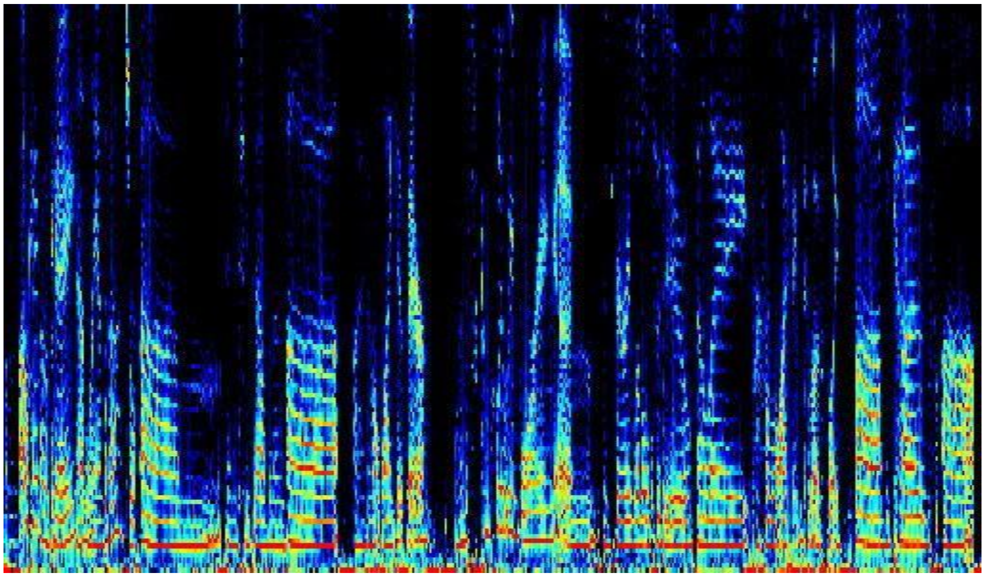

Fig V. Spectrogram representation of a depressed patient's voice

ISSN: 2233-7857 IJFGCN

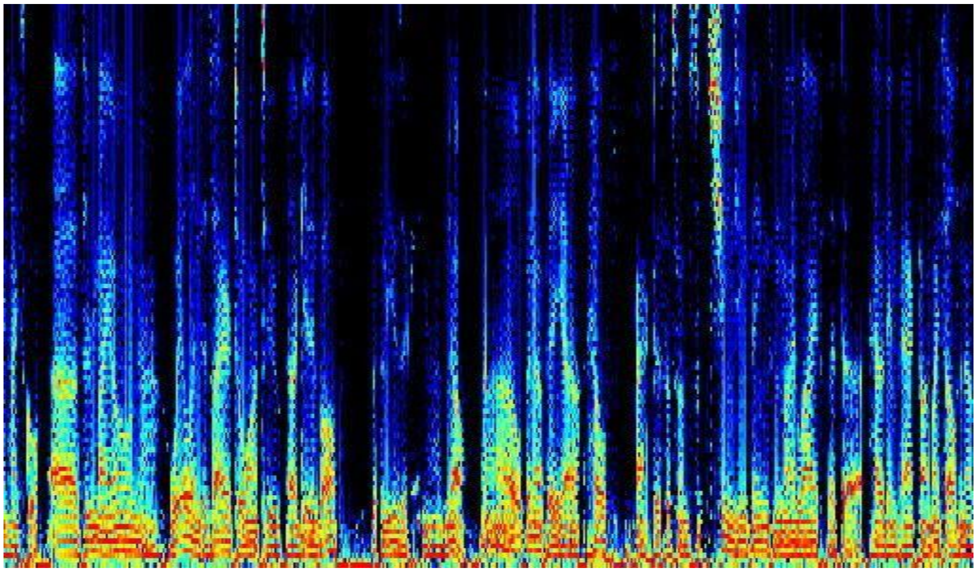

Fig VI. Spectrogram representation of a non-depressed patient’s voice

From Fig V and Fig VI, we can conclude that its not possible for a human to find any patterns in spectrogram of depressed that distinguishes them from the spectrogram of non-depressed patients. Thus, CNN model is trained to find such patterns, which aren't obvious to human eye.

C. Random Sampling

The dataset DAIC-WOZ which was used here had a biasing for non-depressed patients as the number of non-depressed patients is four times more than the number of depressed patients. Thus, if CNN is trained using this dataset, then there will be a non-depressed class bias. If these were used directly, there would be a strong bias in the trained model for non-depressed people. The model might not learn any pattern and just output that the unknown patient is not depressed, which will be a major issue. Additional bias may occur due to the duration of interviews as a large interview of an individual may emphasize certain characteristics that are person specific. To solve this, each participant’s segmented spectrogram was partitioned into 15-second slices after which participants were selected randomly from both classes (depressed and non-depressed) in equal numbers. Then, a fixed number of slices were sampled from each of the selected participants to ensure the CNN has an equal interview duration for each participant. The cropping also ensures that the CNN model gets the input of equal dimensions, which was not possible since the audio length was different for every patient. Thus random sampling is an essential step to have a balanced input to CNN and also minimize the individual effect by random sampling. After the random sampling , the amount of data available was split into training (80%) and testing (20%). The total numbers are shown in TABLE I.

	Depressed	**Non-depressed**
Training Dataset	1003	1002
Test Dataset	251	251

TABLE I. Dataset distribution in training and test data

D. Convolutional Neural Network (CNN) Architecture

The classifier system we used is commonly used machine learning algorithms out there for Image Classification i.e. Convolutional Neural Network(CNN).

Many breakthroughs have been achieved by CNN in the field of image processing and is being used widely for signal processing and speech recognition.

In any CNN architecture, it contains many pairs of convolutional and pooling layers. CNN parameters consist of learnable filters which are applied throughout the input space. A max-pooling layer is used to create a partition in activations of convolution layer to form non-overlapping rectangles and from a sub-region take the maximum activation.

The ideas that make CNN successful are - Local Connectivity and Weight Sharing seen in Fig VII. Local connectivity leads to sparse connections as it restricts a neuron to only connect to its local region and the number of parameters to be learnt is restricted by Weight Sharing. It is common to deploy a pooling layer followed by a convolution layer, replacing the output of the net at a certain location with a summary statistic of the nearby outputs[6]. Pooling is used for down-sampling. Pooling helps the model to be tolerant to minor differences of position of objects. This is useful since we only care about if a specific pattern seen in depressed person is present in the audio or not, and not where it is located.

The architecture used for CNN model is shown in detail in TABLE II. The CNN model consists of 2 convolutional layer (filter size 5*5 and 3*3 respectively), each convolutional layer is followed by a max-pooling layer. These layers were followed by 2 fully-connected layers, each followed by a drop-out layer (0.6 and 0.8 respectively). RELU was used for activation of the nodes, and the output layer had Sigmoid activation. The output size of each layer is shown in the Fig VIII.

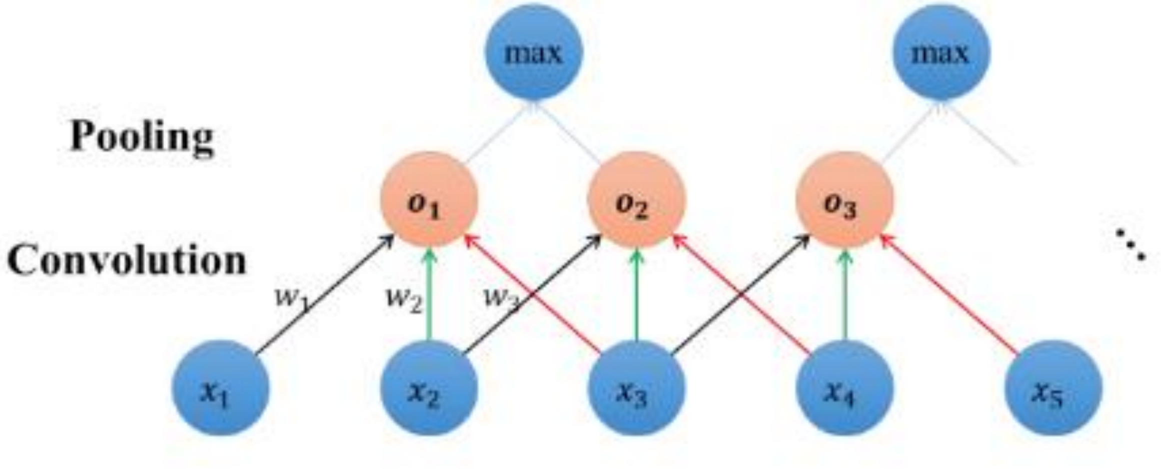

Fig VII. Local connectivity and weight sharing in CNN [2]

ISSN: 2233-7857 IJFGCN

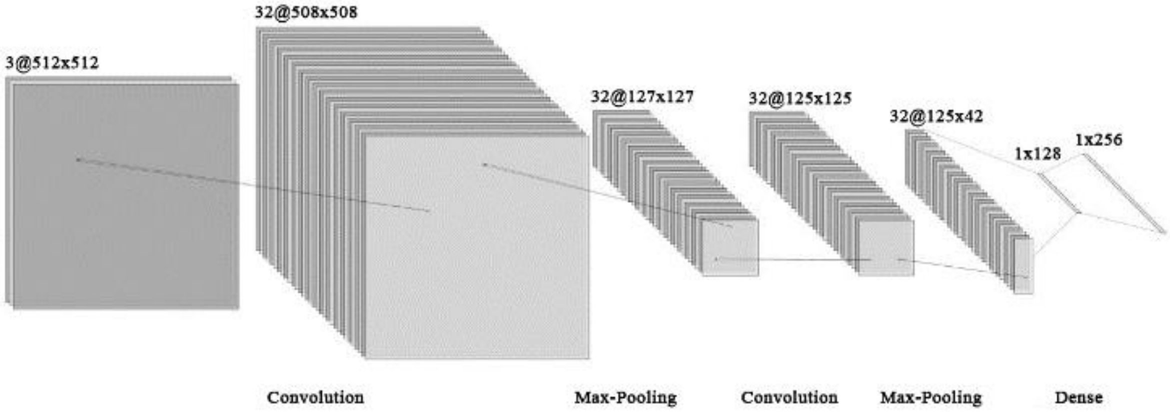

Fig VIII. CNN architecture

Layer No.	Type of layer	Filters
Layer-1	Convolution Layer-1	32 filters (5*5)
Layer-2	Max-pooling Layer-1	(4*4)
Layer-3	Convolution Layer-2	32filter (3*3)
Layer-4	Max-pooling Layer-2	(1*3)
Layer-5	Fully connected layer (256 nodes)-1	-
Layer-6	Fully connected layer (128 nodes)-2	-

TABLE II. Detail architecture of CNN

7. Result and Analysis

The changes in accuracy and loss of the model with respect to number of epochs, is seen in the Fig IX and Fig X respectively. The model keeps increasing the accuracy on the train data as the number of epochs increases, but the accuracy on the test data plateaus around epoch number 40.

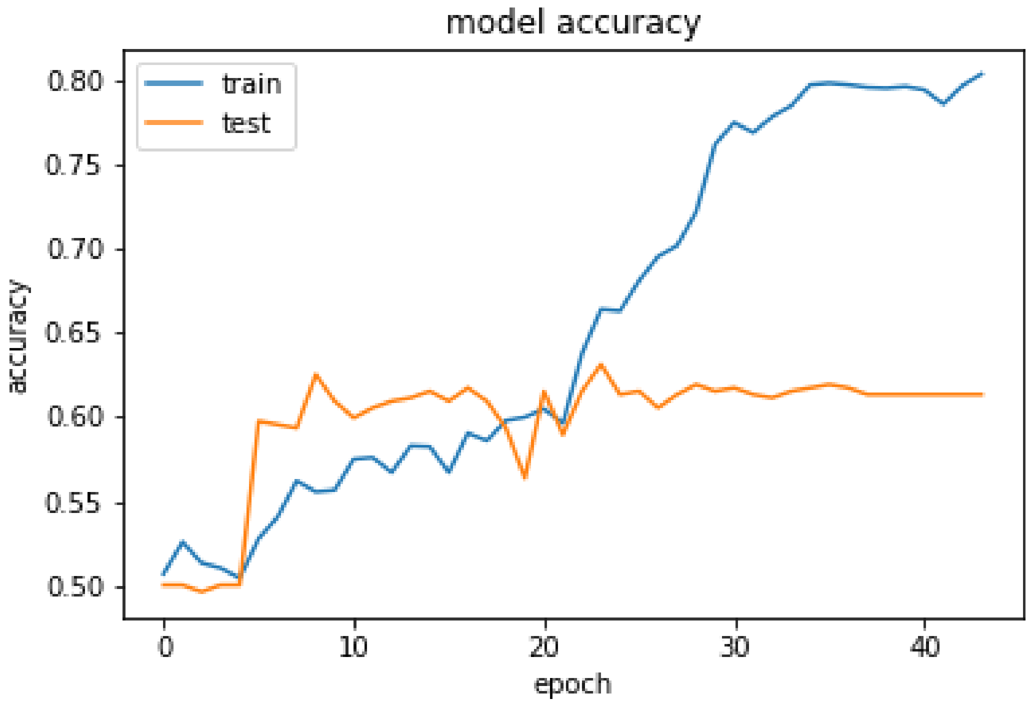

Fig IX. Graph plot of Accuracy vs. Epoch

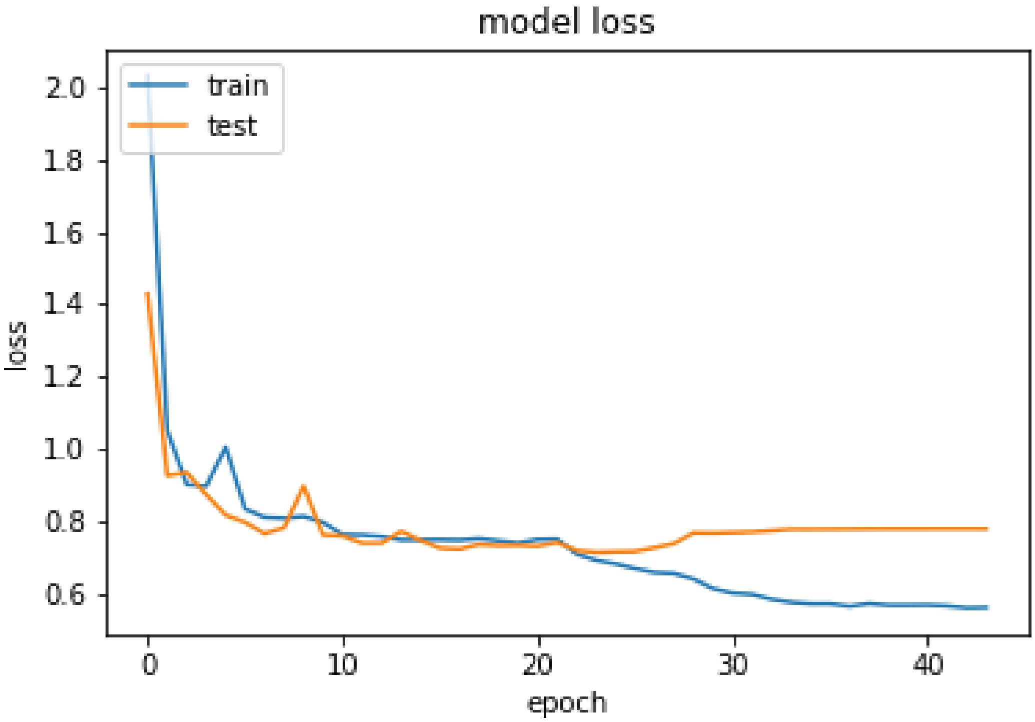

Fig X. Graph plot of Loss vs. Epoch

8. Conclusion

In the paper, We proposed a Deep Neural Network for Automatic Depression detection with speech of the patient. Such Deep learning models find the patterns common in depressed patients and thus help in depression detection for new patients. The trained CNN model has the accuracy of about 0.65. Also, Random Sampling was used to balance both the classes, this removing any bias in the model. The model was trained on a DAIC-WOZ dataset which had a PHQ8 score of patients along with their audio file.

References

[1] H. Long, Z. Guo, X. Wu, B. Hu "Detecting Depression in Speech: Comparison and Combination between Different Speech Type", IEEE International Conference on Bioinformatics and Biomedicine(BIBM), 2017.

[2] X. Ma, H. Yang, Q. Chen, D. Huang, and Y. Wang, "Deepaudionet: An efficient deep model for audio based depression classification", in Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge, ser. AVEC '16. New York, NY, USA: ACM, 2016, pp. 35-42.

[3] R. Shankayi, M. Vali, M. Salimi, M. Malekshahi "Identifying Depressed from healthy cases using processing", Proceedings of the 19th Iranian Conference on Biomedical Engineering (ICBME 2012), Tehran, Iran, 21-22' December 2012.

[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097-1105, 2012.

[5] Y. Bengio. A connectionist approach to speech recognition. International Journal of Pattern Recognition and Artificial Intelligence, 07(04):647-66-, Aug. 1993.

[6] I. G. Y. Bengio and A. Courville. Deep learning. Book in preparation for MIT Press, 2016

[7] "India is the most depressed country in the world", *India Today*, October, 2018

[8] WebMD, 2018

[9] Le Yang, Dongmei Jiang, Ciaohan Xia, Ercheng Pei, Mesiha Cédric Oveneke, Hichem Sahli, "Multimodal Measurement of Depression Using Deep Learning Models", *AVEC' 2017*.

[10] "What are the benefits of Speech Recognition Technology?", *IEEE Signal Processing Society*, 2018.

[11] Tsang-Long Pao; Jun-Heng Yeh; Yao-Wei Tsai, "Recognition and analysis of emotion transition in mandarin speech signal", *IEEE CONFERENCE PUBLICATIONS*, 2010, Page(s): 3326-3332.

[12] P Gangamohan, V.K. Mittal,; B. Yegnanarayana, "A Flexible Analysis Synthesis Tool (FAST) for studying the characteristic features of emotion in speech", *IEEE Conference Publications*, 2012, Page(s): 250-254.

[13] C. Lee, S. Narayanan, and R. Pieraccini, "Recognition of negative emotions from the speech signal", in *IEEE Workshop on Automatic Speech Recognition and Understanding*, 2001, pp. 240-243.

[14] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, C. Votsis, s. Kollias, W. Fellenz, and J.G. Taylor, "Emotion recognition in human-computer interaction", *IEEE Signal Processing Mag.*, pp. 33-80, Jan 2001.

[15] A. Ozdas, R.G. Shiavi, S.E. Silverman, M.K Silverman, and D.M. Wilkes, "Investigation of vocal jitter and glottal flow spectrum as possible cues for depression and near-term suicidal risk", *IEEE Trans.Biomed.Engg.*, vol.51, no 9, pp. 1530-1540, Seo. 2004.

[16] E. Moore, II, M. Clements, J. Piefer, and L. Weisser, "Critical Analysis of the Impact of Glottal Features in the Classification of Clinical Depression in Speech", *IEEE Transactions on biomedical engineering*, Vol. 55, No.1, Jabuary 2008.