Depression Detection

Vidisha Arvind
Department of Computer Science
School of Engineering and Applied Sciences
Bennett University Greater Noida, U.P., India

Abstract— Neurophysiological changes may arise at the moment of detection of depression, which may affect laryngeal function i.e. the actions of the vocal folds. It is an important task to specify these changes in a specific way from a patient's speech signals, as this usually requires accurate isolation of the information from the voice source. We present work to encourage the use of sensing technologies by clinicians with an emphasis on acoustic characteristics from spontaneous speech to depression. The purpose of this experiment is to explore the differences between a depressed and a healthy individual in the positive and negative state of emotions. We have used the 24 Actors interviews in the RAVDESS Dataset for this. As a result, we used the MLP model that gave approximately 87.63 percent precision and the CNN model that gave 58 percent precision.

Index Terms— Speech signal processing, Depression, Neural Network, Multilevel Perceptron, Acoustic features, speech-based depression detection,MFCC(mel frequency cepstral coefficients) ,Chroma.

1 Introduction

Unlike much of the studies that compared depressed to the non-depressed applicants with respect to the intra personal behaviours at one point in time, people could focus on both of the intra and inter personal traits. Acoustic features can also be found as the promising predictors of the depressive symptoms. We have observed and investigated whether a person is being able to recognize the acoustics of the depression or not, and the other quantitative means to observe that extent is to which extent the acoustics of a person can change in the severity over the course of early depression symptoms.[1] To observe whether the vocal prosody of a person varies as the individual recovers from the depression, longitudinal studies can be necessary that assess the changes in depression severity over the period of the early symptoms of depression.

Depression had been associated with the increase in white blood cells and the inflammation in the body. It is said to evolve as a protection mechanisms for our bodies when our bodies were prone to infectious attacks, thus causing inflammation and increasing the temperature and the number of white blood cell, our body sees it as it is protecting ourself. But multiple time it leads to chronic inflammation which in turn leads to chronic depression.[11] We need to build our approach based not on that a depressed person is probably weak and is cringy, but rather the fact that the person may have his body in survival mode and has a very overreactive immune response since his body is in the illusion that it's in danger. The acoustics part may sound a bit out of the topic but it isn't. The way a person voices changes is surprising. The voices of a person get very variable and the number of other changes that we will explore more in this paper.

Prosodic features can be found as promising predictors of depression.[5][12] Acoustics (including prosody) of

a person is helpful in detecting the early symptoms of depression and they changes time to time. We observe and investigate whether the person can perceive vocal acoustics of depression, and quantitative methods to investigate the extent to which the acoustics of a person change in severity over the course of early depression symptoms. To observe whether acoustics varies as an individual recovers from depression, studies are required which assess the change in depression severity over the course of depressive disorder. [4]

2 Literature Work

Depression is known to be a disorder which is characterized by prolonged emotional imbalance.Although, the causes of depression remain unknown at present times and it is quite certain that most of the aspects such as psychological, biological, and even social environment are involved in the pathogenesis of the depression.In the current scenario, the detection of depression is almost depending on doctor-patient communication. During the process of communication between doctor and the patient, depressive patients tried to hide their true feelings because of their own reasons. In addition to that, psychiatrists can use different detection and testing criteria for different number of patients accordingly to their own professional experience. Therefore, to find out the biological indicators that could possibly reflect depression state which has become the open research direction of numerous researches.

For the audio-recognition part, scholars had proposed multiple methods. In one study, they made a 1- dimensional LSTM(long-short term memory) and a 2 dimension LSTM to extract multiple global and local speeches which were related to emotions. This can improve the precision of the native mode by joining the two features. The study extracted the forms(and features) of audio and video and combined them into different signs of abnormal behaviours. After, LSTM RNN(Recurrent Neural Network) was used to show dynamic time period information. The audio-recognition method was used to extract features from MFCC("Mel frequency cepstrum coefficients"). Then, a matrix is used to show the energy band mode defined for the specific frequency and time frame.

2.1 Events in Depression Acoustics

The typically used audio features for sentiment classification these days, including the analysis of depressed and non-depressive populations, are one and only the acoustic features derived at every frequent interval within the range of fixed-length speech frames (20 ms)[5]. For example, MFCC shows the spectral characteristics of an audio signal segment. Frame features provides descriptive content-based details about the audio signals. This information, however, includes some repetition as consecutive frames transmit the same piece of information and can also be too precise because MFCCs bring unique information oriented to the informative phonetic and speaker, which can help to contribute undesirable differences. [14], [5].

2.2 Detection of Depression from voice and Facial Features

M.Naseer , A.Jati , PG.Shivakumar , SN.Chakravarthula , P.Georgieu have discussed about multimodal depression 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[16]. They showed that Teager energy cepstral coefficients (TECC) surpassed standard baseline characteristics in the audio modality, while i-vector modeling 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.

2.3 Detecting Depression with Audio/Text Sequence Modeling of Interviews

T.Alhanai, M.Ghassemi and J.Glass demonstrated their work of an Asynchronous distress detection method that analyzed interviews between a person and the representative that learned from the sequences of questions and answers without the need to precisely model the content of the subject[15]. They used data from 142 people

who have undergone the depression diagnostic test and modeled the interactions in a Long-Short Term Memory (LSTM) . Their results were similar to approaches specifically centered on interview questions that suggested that depression can be characterized by serial modeling of an interactive-tio-sion with minimal information about the essence of the interview..[13]

In order to do detection whether people were depressed or not during their interviews. The three experiments were conducted where audio and text modal qualities were modeled (1) without the question that bring on the responses, (2) with the context by exerting on the question asked, and (3) and with respect to the sequence of the responses.

As a result , it gave bad performance for context free model in text based weighted model and in audio based model the context free model gave better results.

3 Methodology

3.1 Dataset collection

RAVDESS [18] data is comprised of 7356 files. On emotional validity, strength, and genuineness, each file was scored 10 times. There were 247 individuals that were characterized by an un-trained adult study candidates belonging to North America were given scores[13]. Test-retest data was given by a further collection of 72 participants. High emotional validity levels, reliability of interrater, and reliability of test-retest intrarater were recorded.

There are 7356 files in the RAVDESS data of size 24.8 GB. In the database, there are also 24 trained actors (12 male, 12 female), in a North American neutral voice, clearly expressing two linguistically related phrases.

Speech includes expressions of disgutsted, astonished, depressed,fearul,angry,happiness, 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).

The 7356 RAVDESS sound files each have a different names (e.g., 02-01-06-01-01-02-01-12.mp4) consists of the filename.

3.2 Data Cleaning

Noise reduction**:** There are several methoda to remove the noise from a given audio clip. All of it requires is a small sample where there is only a background noise present, and then spontaneously deletes this noise from the rest of the sampled audio.
The steps of the algorithm are :

i. The FFT is determined using the noise audio recording. Statistics are calculated using the noise in frequency over FFT.

ii. Depending on the noise statistics (and the desired algorithm sensitivity), a threshold is determined. (Fig – 1)

iii. Over the signal, FFT is determined

iv. Comparing the signal FFT to the threshold defines a mask

v. The mask is smoothed through frequency and time with a filter (Fig - 2)

vi. The mask is added to the signal's FFT and is reversed.

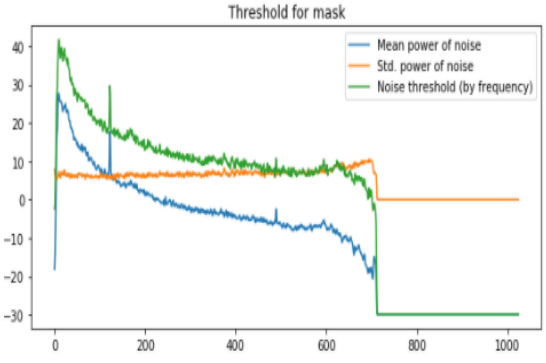


Fig-1 : Threshold calculated over noisy audio.

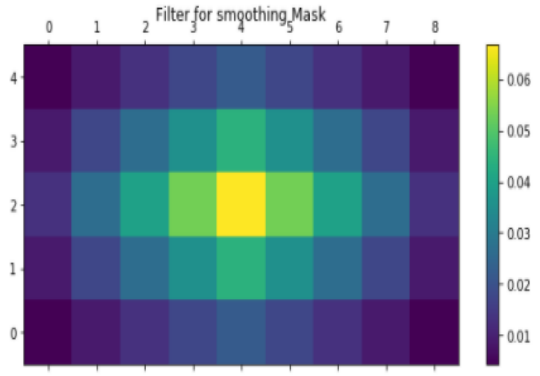


Fig- 2 :Smoothening of mask by applying filter.

3.3 Speech Recognition

Speech recognition is the way of converting acoustics(speech of a person)[4] 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.(Fig-3)

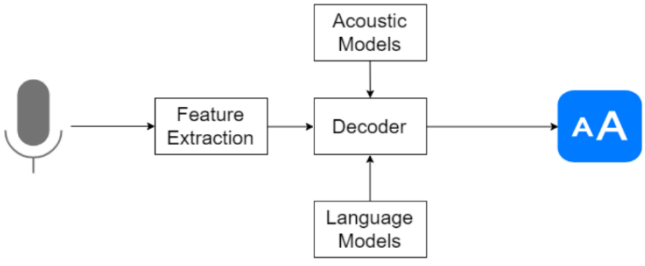


Fig-3 Speech Recognition from Audio file (Python)

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.

3.4 Feature Extraction

Acoustic Features: In general, the more precise and very basic features of audio to recognise 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.

Some of the main audio features considered for this experiment are as follows:[6][8]

MFCC(Mel-Frequency Cepstral Coefficients): 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 toungu, 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.[6] Options are considered for the lower dimensions of the 1st 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 caputes melodic and harmonic characteristics of speech or an audio signal. It is consisting of 2 features:[3]

a) Chroma Vector: It has twelve element expression of spectral energy.
b) Chroma deviation: It is the twelve chroma parameters standard deviation.

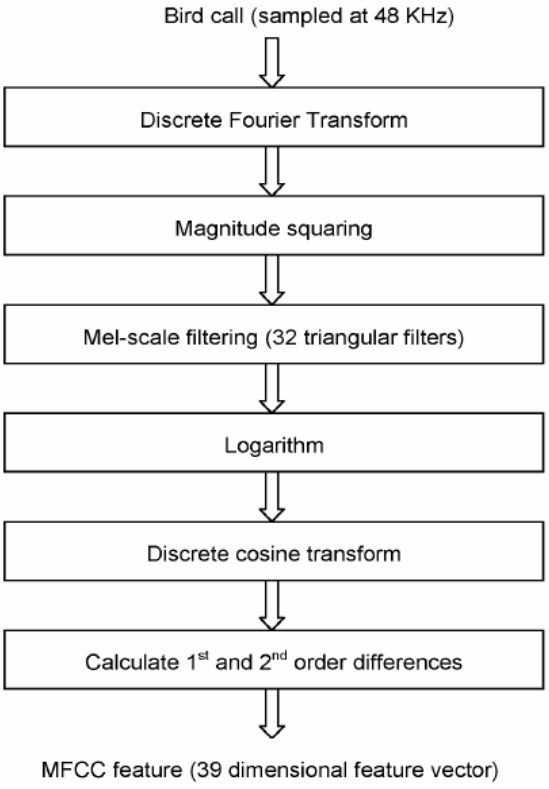


Fig-4 MEL frequency extraction

3.5 Convolutional Neural Network

The deep learning model depending upon the convolutionary neural network (CNN) is used and its dense layers[9] have been used. As the only audio feature to train our cnn model, the MFCC and chroma fea-tures 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 [6], 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 audiotory 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 < set of training samples > x n x 1 on which we executed a one-dimensional CNN round as the activation function ReLu (Fig-5),[12] and 2 x 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. Pooling allows the cnn model to focus only on the main characteristics of each of the data components, not segregating them by their position.

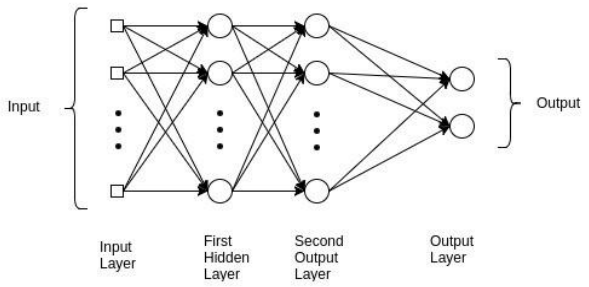
Layer (type)	Output Shape	Param #
conv1d_40 (Conv1D)	(None, 65, 256)	1536
activation_50 (Activation)	(None, 65, 256)	0
conv1d_41 (Conv1D)	(None, 65, 128)	163968
activation_51 (Activation)	(None, 65, 128)	0
dropout_10 (Dropout)	(None, 65, 128)	0
max_pooling1d_10 (MaxPooling	(None, 8, 128)	0
conv1d_42 (Conv1D)	(None, 8, 128)	82048
activation_52 (Activation)	(None, 8, 128)	0
conv1d_43 (Conv1D)	(None, 8, 128)	82048
activation_53 (Activation)	(None, 8, 128)	0
flatten_10 (Flatten)	(None, 1024)	0
dense_10 (Dense)	(None, 7)	7175
activation_54 (Activation)	(None, 7)	0

Total params: 336,775
Trainable params: 336,775
Non-trainable params: 0

Fig-5 Detailed structure of the classifier

3.6 Multi layer Perceptron

Multi layered perceptron is a part of Artificial Neural Networks(ANN). Multi layered perceptron involves 3 layers of nodes atleast in neural networks: output layer, hidden layer and input layer. Excluding our input nodes, every nodes are the neurons which use a non-linear function for activation. MLP activates a supervised learning practice. The techniques is widely known as back propagation in training.[3][7] Its multiple layers and the non-linear activation functions can differentiate between Multi layer perceptron from the other linear perceptrons. It can differentiate the data which is not separable linearly.



There are several single linear layers (combos of neurons) within the Multilayer perceptrons[1]. If we take the simple three-layer network example, the last layer is called the output layer and the first layer is called the input layer , and the hidden layer is called the middle layer. The input data can then be fed to the input layer and the output can be retrieved from the output layer. The number of hidden layers can be raised according to the need, and for the mission to make the model more diverse. (Fig-1)

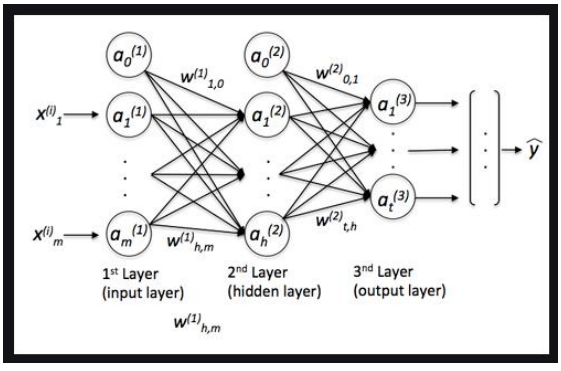


Fig- 6: MLP Architecture

	precision	recall	f1-score	support
Depressed	0.83	0.88	0.85	154

Fig- 7 Accuracy score

5 RESULTS

The findings attained from the evaluation process indicate the efficacy of the model on the RAVDESS 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 called depression in Fig-7. These findings suggest that recall and accuracy are kind of balanced, enabling us to achieve a 0.85 F1 score for the depression class. The slight shift in F1 highlights the robustness of the MLP model, which manages 88.67 percent accuracy effectively. The only ones in which the model is not correct are the category "happy" and "neutral". Over the many research studies, it is accepted that they are the most difficult condition of emotions to discern not only from a person's expression, but also when studying facial movements or analyzing text.

6 CONCLUSION

We tried to create a model architecture based on neural networks for the diagnosis of depression and using the audio clips of the (RAVDESS) actors. But our depression detection model is also trained in the person's speech to distinguish seven distinct states of emotions (happy, depressed, angry, terrified, indifferent, relaxed, disgust, surprised) and we have obtained an overall Accuracy score of 0.87 with the best performance on the depressed class category 0.88 and the worst performance on the happy class 0.81.

We have extracted the MFCC,MEL and Chromagram 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. But our final choice of our deep learning model was the MLP that on the test set attained an accuracy performance of 0.87.

The result gained can only be worth it as a starting point for further expansions, updates, and enhancements.

7 LIMITATIONS AND FUTURE WORK

Limitations:

I. Insufficient amount of DATASET.
II. Model performed better for other emotions except joy and neutral.

Future Enhacements: Emotion detection using audio can be integrated with facial emotion recognition that will help achieve better accuracies.

References

[1] Zogan, Hamad, Jameel Shoaib, Wang Xianzhi, and Xu Guandong. Depression Identification of Multi-Modalities using a Hybrid Deep Learning Model on Social Media. Preprint arXiv arXiv:2007.02847 (2020).

[2] Xezonaki, Danai, Ale-xandros Potamianos, Georgios Paraskevopoulos, and Shrikanth Narayanan. Applied to Depression Detection from Transcribed Clinical Interviews, 'Effective Conditioning on Hierarchical Networks.' arXiv preprint arXiv:2006.08336 (2020).

[3] Chiu, Chun Yueh, Hsien Yuan Lane, Jia Ling Koh, and Arbee LP Chen. "Intagram multimodal depression detection considering time interval of posts." Journal of Intelligent Information Systems (2020).

[4] Anastasia, Ushakova. "Depression Detection from Social Media Profiles." In Data Analytics and Data Intensive Domains Management: 21st International Conference, DAMDID/RCDL 2019, Kazan, Russia, 15-18 October 2019, Updated Selected Papers, vol. 1223, 181 p. Nature from Springer, 2020.

[5] Zhaocheng Huang, Julien Epps, Dale Joachim, Vidhyasaharan Sethu, IEEE Journal of Selected Topics in Signal Processing, 2019 Natural Language Processing Methods in Speech-based Depression Detection for Acoustic and Landmark Event-based Features.

[6] Adrian David Cheok, Owen Noel Newton Fernando, Hooman Aghaebrahimi Samani "An affective interactive audio interface for Lovotics" , Computers in Entertainment, 2011.

[7] Wang, Xinyu, Chunhong Zhang, Yang Ji, Li Sun, Zhana Bao and Leijia Wu. "An anxiety detection model based on sentiment interpretation in microblog social network." In the Information Discovery and Data Mining Pacific-Asia Conference, pp. 201-213. Springer, Heidelberg, Berlin, 2013.

[8] Alghowinem, Sharifa, Roland Goecke, Michael Wagner, Michael Breakspear, and Gordon Parker. "Eye movement analysis for depression detection." 2013 IEEE International Image Processing Conference, pp. 4220-4224. 2013, IEEE.

[9] Evins, Grace G., James P. Theofrastous, and L. Galvin Shelley.' Postpartum depression: comparison of screening and routine clinical assessment.' American obstetrics and gynecology journal 182, no. 5 (2000): 1080-10822-ogy 182, no. 5 (2000).

[10] Woo, Woontack, Park Jong-Il, and Iwadate Yuichi. "Emotion observation from dance performance utilizing time-delay neural networks." In Proceedings of the Fifth Joint Conference on Information Sciences, JCIS 2000, pp. 374-377. In 2000.

[11] McGilloway, Sinéad, Roddy Cowie, Ellen Douglas-Cowie, Machiel Westerdijk, Stan Gielen, and Sybert Stroeve, respectively. "Approaching systematic recognition of sentiment from sound: A basic benchmark." Voice and Emotion in the ISCA Tutorial and Analysis Workshop (ITRW) I n 2000.

[12] [1912.10458] Emotion Recognition from Speech (arxiv.org)

[13] Sharifa Mohammed Alghowinem, Tom Gedeon, Roland Goecke, Gordon Parker, Jeffrey Cohn. Depression Detection Models Interpretation through Feature Selection Methods, IEEE Affective Computing Transactions, 2020

[14] Wymer, Joy H., Randee L. Booksh, and Linda S. Lindman. Aprosody's neuropsychological perspective: characteristics, operation, evaluation and treatment."Applied neuropsychology 9, no. 1 (2002).

[15] Tuka, Mohammad M. Ghassemi, Al Hanai, and James R. Glass. "Detecting Depression with Audio/Text Pattern Modeling of Interviews." in Interspeech, pp. 1716-1720. I n 2018.

[16] Md Nasir, Arindam Jati, Prashanth Gurunath-Shivakumar, Pa-nayiotis Georgiou, Sandeep Nallan Chakravarthula. Proceedings of the 6th International Audio/Visual Emotion Challenge Workshop - AVEC '16,2016' 'Multimodal and Multiresolution Depression Identification from Facial and Voice Landmark Features'.

[17] www.isca-speech.org

[18] www.zenodo.org