

***A Mini-Project Report On***

**“Depression Analysis using Audio Analysis”**

***Submitted By***

**Mangirish M. Bhagat Amonkar**

**Roll No: 06**

**F.Y. M.Sc. (Data Science and Big Data Analytics)**

**School of Computer Science**

**Faculty of Science**

**MIT – World Peace University**

**Pune - 411038**

**Academic Year 2019-2020**

**APRIL – 2020**

**MIT WORLD PEACE UNIVERSITY, PUNE**

**SCHOOL OF COMPUTER SCIENCE**

***Certificate***

This is to certify that

**Mangirish M. Bhagat Amonkar, Roll No. 06,**

Of ***M.Sc. (Data Science and Big Data Analytics)*** successfully completed his/her Mini-Project in

**“Depression Analysis using Audio Analysis”**

to our satisfaction and submitted the same during the academic year 2019- 2020 towards the partial fulfillment of degree of **Master of Science in Data Science and Big Data Analytics** of MIT World Peace University under the School of Computer Science, MIT WPU, Pune.

|  |  |  |
| --- | --- | --- |
| **Prof. Dr. Shubhalaxmi Joshi** | **Prof. Sachin Bhoite** | **Prof. Supriya Aras** |
| **Associate Dean** | **Program Head** | **Assistant Professor** |
| **Faculty of Science** | **School of Computer Science** | **School of Computer Science** |
| **MITWPU** | **MIT WPU** | **MIT WPU** |

**ACKNOWLEDGEMENT**

Presentation, Motivation and Inspiration have always played a key role in the success of any project. The success and final outcome of this project required a lot of guidance and assistance. I am extremely fortunate to have got this all along the completion of our project work.

I express my sincere thanks to my mentor guides, Prof. Prajakta Soman and Prof. Supriya Aras. Also, I would like to thank my fellow mates who helped incase of any queries asked.

I pay my deep gratitude to everyone of them to encourage me to the peak and help me throughout the project. It is because of your valuable guidance and kind supervision throughout the course of the project that helped me to complete this project.

Mangirish M. Bhagat Amonkar

Roll No. 06

Contents

[Introduction 5](#_Toc52214906)

[Domain Name 5](#_Toc52214907)

[Motivation 5](#_Toc52214908)

[Problem Statement 5](#_Toc52214909)

[Literature Survey 7](#_Toc52214910)

[Solution Design 8](#_Toc52214911)

[Solution Approach 8](#_Toc52214912)

[Technology Stack 8](#_Toc52214913)

[Design Model 8](#_Toc52214914)

[Solution Implementation and Results 9](#_Toc52214915)

[Obtaining Data 9](#_Toc52214916)

[EDA 9](#_Toc52214917)

[Pre-Processing 9](#_Toc52214918)

[Machine Learning Algorithms Used 9](#_Toc52214919)

[Results 9](#_Toc52214920)

[Conclusion and Future Work 10](#_Toc52214921)

[Conclusion 10](#_Toc52214922)

[Future Work 10](#_Toc52214923)

[References 11](#_Toc52214924)

# Introduction

## Depression Detection using Audio Feature

Depression is most reported common mental disorder. Individuals suffering from depression often diagnosed with social anxiety, major depressive disorder and other addictions also, which is commonly known as comorbid anxiety and disorders. It can impact every area of your life, including how you sleep and eat, your education and career, your relationships and health. Depression doesn’t just occur for an individual in a vacuum, it can affect your friends, family, co-workers, and everyone around you. Besides, depression may impact how you perform at work or your levels of concentration, so it can negatively affect productivity. Leaving depression untreated can lead to many other complications in one’s personal and professional life. This is why it is so important to seek out help, not just for relationships and work, but for your own sake.

Many aspects are their which can be taken in consideration to identify depressed person, most common are talking behaviour and voice. Before declaring the person is depressed psychiatrist interview, try to know the person and also take some tests. Then the person is declared clinically depressed. We thought of working with audio data and in this project we are using spectrograms which is visual representation image of audio’s frequency.

## Motivation

As we know the Prime Minister declared lockdown in mid of March, everyone’s routine almost has changed. Home became the new environment for many of us and reduced physical activities. Gadgets became our new friends as meeting our real friends wasn’t possible. This sudden change affected human mental state and their lives. We came across news articles and incidents which were happening because of this transformation. After reading many articles in the newspaper we had shallow research on the “Mental Health” which was impacted majorly. We got an idea of detecting depression at early stage would help many lives and families. So, we thought of using Deep Learning to make this happen and started researching. This led us to the amazing research by the USC’s Institute of Creative Technologies’ Prof Jill Boberg and her associates.

## Problem Statement

Seeing world going through major transformation we perceived that depression is being common disorder among us and identification of it is important. Speech is one of the factors to detect a person as depressed. In this project we are trying to detect depression using spectrogram image- the visual representation of speech frequency. As a solution we are training a Convolutional Neural Network(CNN) model with spectrogram images of both non-depressed and clinically declared depressed people, which will result in detecting new spectrograms images as non-depressed or depressed.

# Literature Survey

As, some of us were going through news material containing depression related articles provoked us to do some part of work related to it. We started researching about previous and ongoing work associated to this domain. We came across to one good link to the research done out in 2014 called DAIC-WOZ for mental health where many aspects of mental disorders like PTSD, anxiety, stress, depression were studied. To study these types of disorders a periodic population-based surveys are conducted. A popular measure to assess is the Patient Health Questionnaire nine-item depression scale (PHQ-9) is been used since 2002, which is further drilled down to PHQ-8 since 2006. We are studying a small module of PHQ-8 that is **Depressive Symptoms** collected from an audio-video form.

We decided to work with audio part using its visual image. This whole concept of audio recognition and its visual images was new concept for us so we started studying about audio analysis and its physics part. Studying the audio acoustic features was necessary and we did that by going through many websites and research papers. Applying this studied concept to create a solution using Deep Learning was next phase to be understood.

We explored audio supported libraries in python, as we planned to build a CNN model using Python Programming language. We gone through documentation of many libraries to deeply understand how it can be used in our project. It includes Librosa, Pydub, Pyaudioanalysis and Spectrust.

We are using spectrogram as images data to feed into CNN model. Spectrogram images are visual representation of the spectrum of frequencies of a signal as it varies with time, which are usually different from normal images. Understanding these images was also a part of our literature work.

# Solution Design

## Technology Stack

Hardware

* RAM - 16 GB
* CPU – Intel i7
* GPU –Nvidia 1650

Software

* Colab
* Jupyter notebook
* Google drive for Data Storage
* MS-Excel

Libraries

* pyAudioAnalysis
* librosa
* pydub
* hmmlearn
* eyeD3
* spectrust
* google.colab
* drive
* scipy.io.wav
* wave
* audioBasicIO
* audioSegemtation
* math
* AudioSegment
* numpy
* cv2
* tensorflow
* keras
* callback
* layers
* regularizars
* preprocessing
* matplotlib
* Sequential
* Con2D
* Max2D
* Dropout
* Flatten
* Dense
* BatchNormalization
* ImageDataGenerator
* EarlyStopping
* IPython.display
* Sklearn.metric

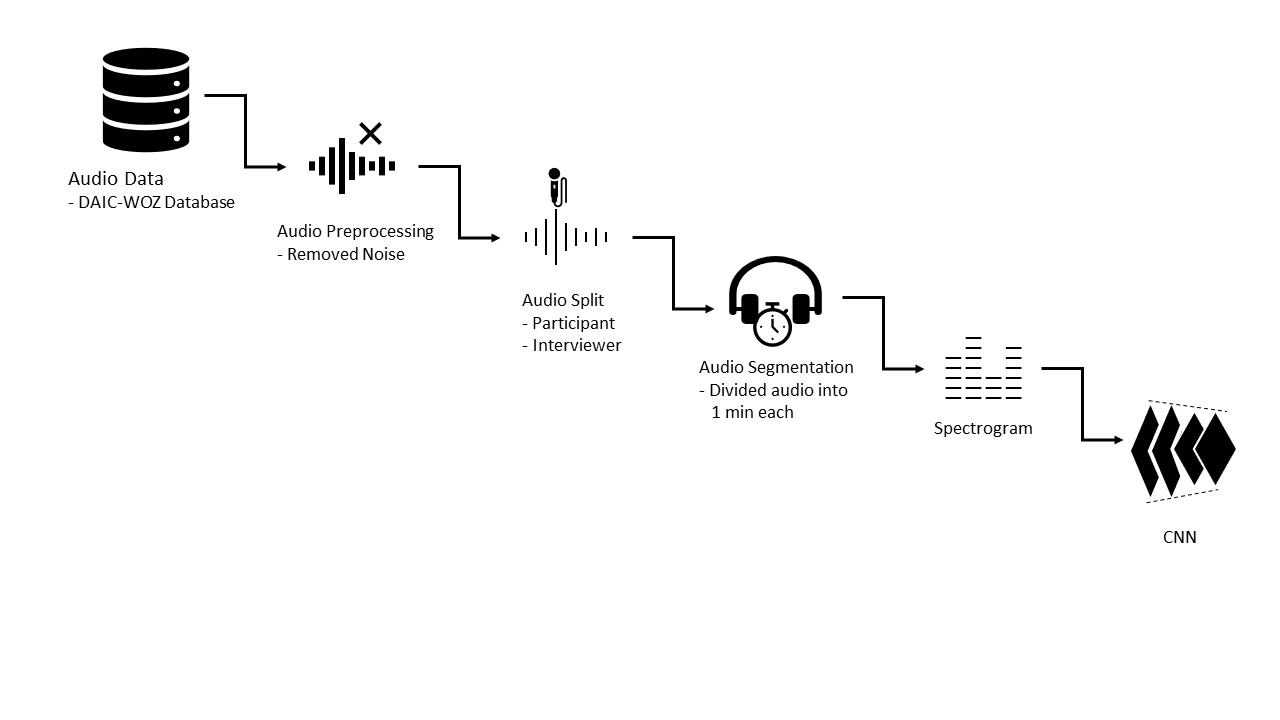
## Design Model

Major part while designing this project was on the audio processing and feature extraction, and rest was the training the model on convolution neural networks. Model was design on waterfall model, which follows the same path throughout the project.

Raw audio data was collected and stored on Google Drive. Using audio analysis libraries and functions we removed noise and preprocessed the data. As only participant voice is required for analysis, we split the voice of participant from interviewer and created a new audio file. These files were further sampled in a minute division for better feature extraction and training of model.

Feature extraction was done using spectrogram. Spectrograms were created on sampled audio files using spectrust library. These spectrograms were labeled in two categories depressed and nondepressed.

Spectrograms were passed to CNN model for training. CNN model ran for 10 epochs and accuracy, precision and recall were noted.



# Solution Implementation and Results

## Obtaining Data

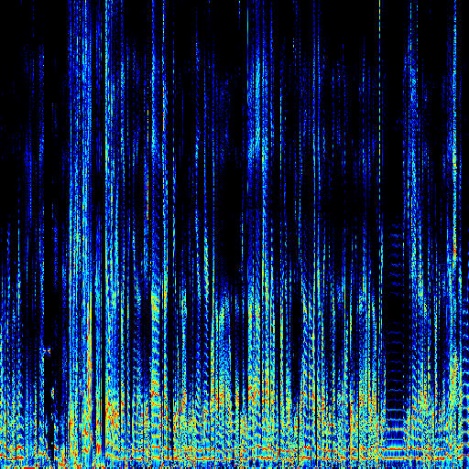
We get the data from the USC Institute of Creative Technologies from Prof Jill Boberg from her research in The Distress Analysis Interview Corpus of human and computer interviews. They carried a series of the interview out by human-controlled robot called Ellie. She asks various questions from basic routine to problems that a person faces. A voice recorder catches the decent frequencies of both interviewer and participant. They interviewed a total of 186 persons in the process out of which 58 people were depressed and the rest were non-depressed. After contacting Prof Jill Boberg, she handed us huge data of roughly 100 gigabytes which divided into 186 folders. Each folder contains around 10 files containing important audio and visual data- .wav file, five files of gesture data collected by video technology like features, features in 3D, gaze, pose, etc. one covarep, transcript, and formant file. As per our planning we were supposed to use audio files so, \*.wav file was saved while others were backed up for further studies.

## Pre-Processing

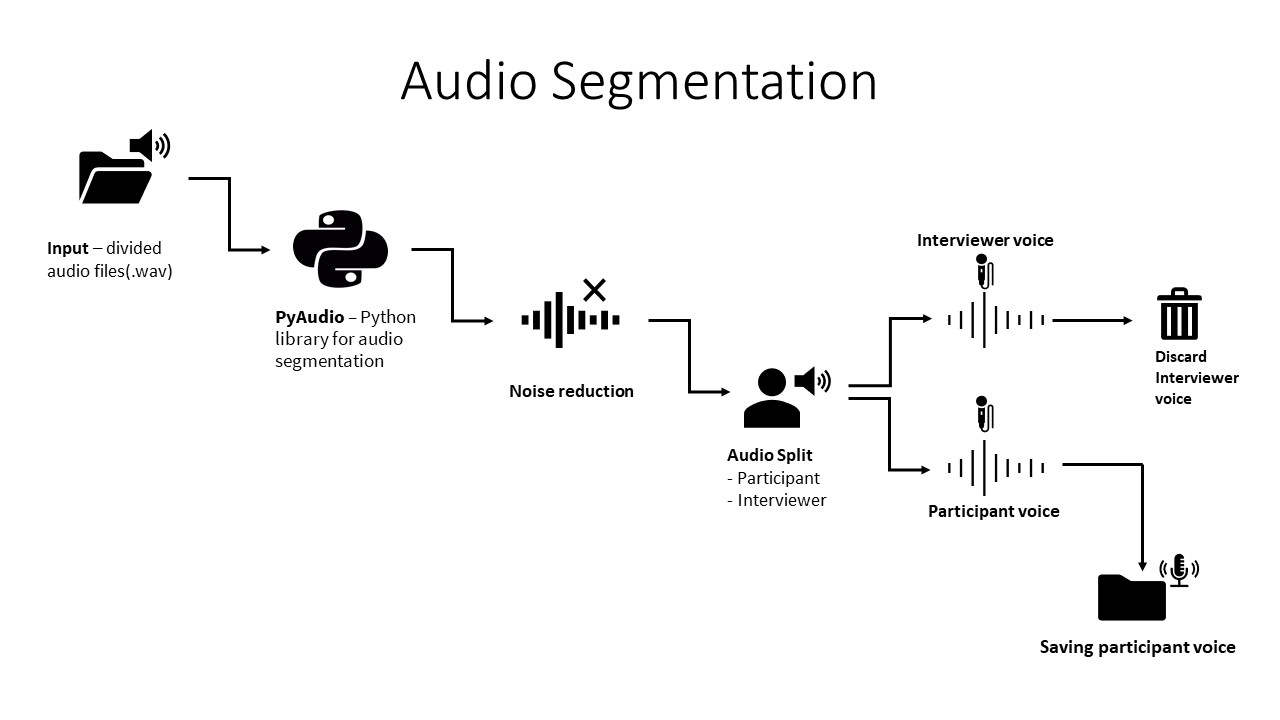
## Since the data was in audio format and raw, we processed the data using library pyAudioAnalysis for removing the noise from every file we received. For feature extraction, we only needed the participant’s voice. Hence, we split the voice from Ellie and create the new audio file i.e. \_no\_silence.wav which only contains the voice of a participant. As the data is small for the Deep Learning approach, there was a need to increase the data. Without disturbing the corpus, we decided to further split the participant file into samples audio files. In this approach, we successfully increased the data points. We created the samples of 1 minute each. This way we created 1000+ files for our model. Yet the data was imbalanced so we used the concept of oversampling in which we copied some random files to equal the images of both depressed and non-depressed category.

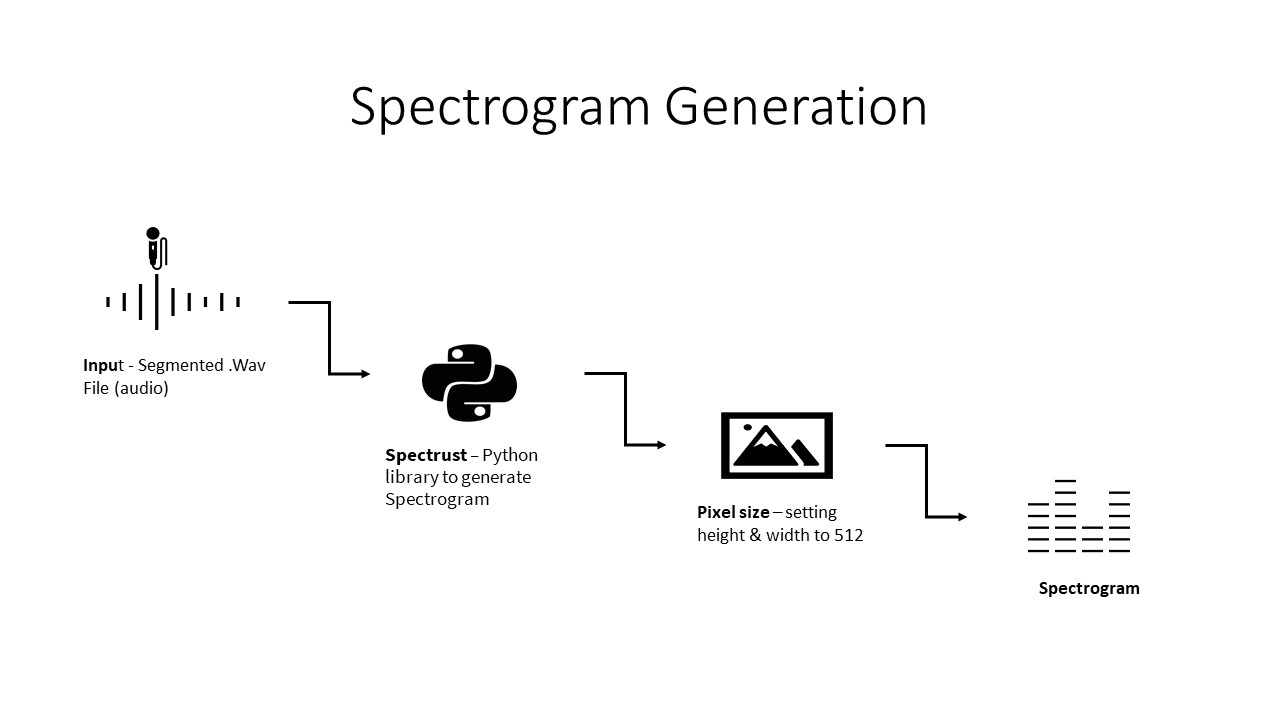
## Feature Extraction

According to our planned approach we are using CNN, and Deep Learning CNN models usually trains on the images. The visual image of audio is what we are using here for that we need to convert the audio files into images. An approach for converting audio into image is called Spectrogram. Spectrogram is a visual representation of audio in image. It is graph of amplitude of frequency of signal over time. Spectrograms were created using library spectrust. These spectrograms are representation of sampled audio files. These images were labelled and split in train, test for further training of model.



Below is the flow of Pre-processing part-





## Deep Learning Algorithms Used

In Deep Learning, a convolutional neural network (CNN, or ConvNet) is a class of deep neural networks, most commonly applied to analyzing visual imagery(Wikipedia). In the case of the spectrogram we feed input as a representation, with the intensity representative of the audio specific frequency with time. CNN begins by learning features like vertical lines, but in subsequent layers, begins to pick up on features like the shape of the frequency-time curve. Such learned features may provide an elegant and powerful representation of different prosodic features of speech, which in turn are representative of underlying differences between depressed and non-depressed speech.

Convolutional Neural Network Architecture-

Input Data- Spectrograms images which were of 512x512 dimension, RGB images and representing time stamp of 1 minute and audio frequency. Rescaled the input data by dividing it with 255.

Data Augmentation-We are using ImageDataGenerator for real time data augmentation. Generate batches of tensor image data with real-time data augmentation.

CNN used here begins with an input layer being convolved with 32-3x3 filters to create 32 feature maps followed by a ReLU activation function. Next, the feature maps undergo dimensionality reduction with a max-pooling layer, which uses a 2x2 filter with a stride of 2.

A second convolutional layer is employed with 128-3x3 filters followed by a max-pooling layer with a 2x2 filter. After this, the flatten layer is added as flattening transforms a two-dimensional matrix of features into a vector that can be fed into a fully connected neural network classifier.

After flatten, the dense layers are connected. The first dense layer is added with 128 neurons, a dropout layer of 0.5 is used (meaning each neuron in the second dense layer has a 50% chance of turning off after each batch update), with L2 regularization. Added one more dense layer with 32 neuron and L2 regularizer. The activation function ‘Relu’ was same for both the dense layer. Lastly, a sigmoid function is applied, which returns the probability that a spectrogram is in the depressed class or nondepressed class. The sum probabilities of each class are equal to 1. A batch size of 10 was used along with an SGD optimizer, which dynamically adapts the learning rate based on the gradient.

While running the model the data after every epoch was being shuffled by keeping shuffle=true. Also, our model is using Callback function monitoring the valid data accuracy with patience=5.

## Results

* CNN model ran for 10 epochs and resulted with 95% of training accuracy. Test accuracy certainly drops to 64%. Accuracy was obtained by early stopping method.
* Other metrics like precision and recall were 82% and 55% respectively which were quite satisfactory.

# Conclusion and Future Work

## Conclusion

We successfully predicted 64% of test spectrogram images to the concerned category. Training with large amount of spectrogram images will make our model more robust, which can lead to make this model get into real-time use.

## Future Work

## State of the emotion detection models exhibit with high precision, but they are NLP based model. As is this model is not yet at a predictive state for practical usage, but rapid development suggests a promising new direction for using spectrograms in depression detection. The main future goal is to create an end-to-end model with the use of IoT devices which can help humans to detect early signs of depression and the severity of the depression.

# References

List the references in IEEE format