Automated Depression Detection using Audio Features

Suraj G. Shinde[1], Atul C. Tambe[2], Avakash Vishwakarma[3], Sonali N. Mhatre[4]

[1]Student of Information Technology Dept., Bharati Vidyapeeth college of Engineering, Navi Mumbai
[2]Student of Information Technology Dept., Bharati Vidyapeeth college of Engineering, Navi Mumbai
[3]Student of Information Technology Dept., Bharati Vidyapeeth college of Engineering, Navi Mumbai
[4] Assistant Professor of Information Technology Dept., Bharati Vidyapeeth college of Engineering, Navi Mumbai

---

Abstract - *Depression is among the foremost common and harmful mental health issues that are rapidly affecting lives worldwide having an enormous impact on well-being and functionality, and important personal, family, and societal effects. Depression not only affects emotional state, but also the physical state of the person. Lacking objective clinical depression assessment methods is the key reason that several depressive patients can't be treated properly. Automatic depression assessment supported visual signals could be a rapidly grown research domain. Human being's cognitive system is often simulated by artificially intelligent systems. Much recent research has shown a relationship between mood changes and changed emotional attention. The analysis of audio features can be an important tool to help in depression identification and monitoring over the course of depressive disorder and recovery. In such a low mood, the voice of human beings appears different from the ones in normal states. In this, an artificially intelligent system is proposed to identify depression using audio.*

Key Words: **AVEC 2014, Depression, Depression Detection, Neural Network, Vocal Expression, Deep Learning.**

1. INTRODUCTION

In recent times, the study on mental health problems has been given growing recognition from various fields in modern society. Depression is a disorder that involves a persistent sense of sadness and loss of interest. It is different from the mood changes that people regularly experience as a part of life. The symptoms of depression can include a loss of desire, changes in appetite, unintentional weight loss or gain, sleeping an excessive amount of or too limited, anxiety, restlessness, slowed movement and speech, loss of energy, and feelings of guilt, etc. Majority of the people that obtain therapy for depression do not recover from it. The illness remains with the person within the sort of insomnia, excessive sleep, and fatigue, loss of energy or digestive problems During a phase of Depression, a person doesn't feel comfortable talking to others about his/her problems with anyone. So, the mental health of human beings is disturbed sometimes that cannot be identified by a normal person.

Among all psychiatric disorders, major depressive disorder (MDD) commonly occurs and heavily threatens the mental health of human beings. 7.5% of all people with disabilities suffer from depression, making it the largest contributor, exceeding 300M people. In recent times, the study on mental health problems has been given growing recognition from various fields in modern society. According to the World Health Organization (WHO), 350 million individuals are suffering from depression, globally.

Artificial intelligence methods can be used for depression analysis with machine-based automatic early detection and recognition to eventually reduce its potential harm in real life. The mental health issues can benefit from these techniques, as they understand the importance of obtaining detailed information to identify the various psychiatric disorders.

2. RELATED WORK

In depression analysis, Cohn et al. [2], who is a guide in the affective computing area, performed research where he fused both the visual and audio characteristics to include behavioral considerations, which are heavily related to mental disorders. Their findings suggest that building an automatic depression recognition system is possible, which will benefit clinical theory and practice. Yang et al. explored variations in the audio features of participants and found modest predictability of the depression scores based on a combination of F0 and switching pauses.[2]

Jain et al. [4] introduced using the Fisher Vector (FV) to encode the LBP-TOP and Dense Trajectories visual prosody, and LLD audio prosody. Cannizzaro et al. [11] found that a change in the severity of depression covaried with vocal prosody. The possible interpersonal impact has been omitted. Depression is heavily associated with personal differences in neuroticism, introversion, and conscientiousness [12]. To investigate whether audio features vary as individuals overcome from depression, studies are needed that evaluate the change in depression severity throughout the depressive disorder.

The few that exist [6], [2], [4] recommend that vocal timing and F0 may be responsive to healing from depression. Kuny and Stassen [3] and Alpert et al [1] observed that intrapersonal gap span and speaking rate are closely associated to change in depression severity. With one exception [3], however, important studies have been limited to inpatient samples that are more critically depressed than those found in the community. They also manage to use structured speaking tasks, which leave open the question of whether audio features in depression impacts interviewers and turn-taking, which is known to affect agreement [7]. The patterns of change in audio and visual expressions have been utilized for automated, contact-free analysis and analysis of depressive behaviors. Some researchers have revealed the characteristics of voice and speech change in depressed individuals.[10] Jan et al. combined features extracted from facial expressions using a deep learning model, and vocal expressions using regression techniques.[11]

3. System Methodology

Human voices and acoustics expressions in depression are theoretically different from those under normal mental states. An attempt to find a solution for depression scale prediction is achieved by combining dynamic descriptions within vocal expressions.

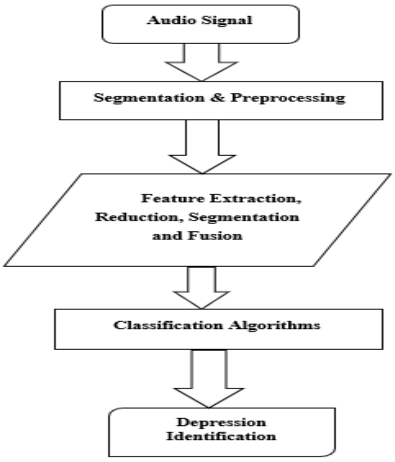
Audio Signal
Segmentation & Preprocessing
Feature Extraction, Reduction, Segmentation and Fusion
Classification Algorithms
Depression Identification

Fig -1: The proposed automated depression detection using audio features

A. Dataset Evaluation

To evaluate the effectiveness the proposed system, we carry out extensive experiments on the DAIC-WOZ dataset, in the Depression Sub-Challenge at AVEC 2014. The dataset, experimental protocols, and prediction results are introduced in the following subsections.

1) Dataset

DAIC-WOZ is part of a large corpus, namely the Distress Analysis Interview Corpus (DAIC), which contains clinical interviews designed to support the diagnosis of psychological distress conditions, such as anxiety, depression, and post-traumatic stress disorder. The interviews are collected by a computer agent that interacts with people and identifies verbal and non-verbal indicators of mental illness. This collection includes audio and video recordings and extensive questionnaire responses, where the part of the corpus contains the Wizard-of-Oz interviews. Samples are reproduced and interpreted for a variety of verbal and non-verbal features.

2) Data-preprocessing

The audio files were cut programmatically using Python and Sox into 15 second samples starting from the 60th second of each recording to capture speech samples, rather than background noise. The length of each recording was normalized to 15 s and it turned out to be satisfactory. We wanted to capture a few (2–3) sentences, so that our method could identify key characteristics of speech in the training data, and later match them in the test data. This gave us voice samples of each person that we used in spectrogram generation, succeeded by CNN training and validation. We refer to this original dataset of 107 speech samples as Dataset A. Another dataset was created to check how the method would work on a much larger dataset. The extended Dataset B was created by extracting each 15 s of speech starting from the 60th second of each recording, until the 7th min (the shortest session in DAIC is 7 min). This method produced 2568 speech samples, 720 of each were of 30 depressive problems. The training set of the AVEC'14 consists of an imbalanced number of depressed and not-depressed samples, as shown in Table 1. We noticed that such an imbalanced dataset may limit recognition performance and cause overfitting. Thus, in the first step we re-sampled the dataset to obtain larger-scale data.

Dataset A		Dataset B	
Category	Samples	Category	Samples
Depressed	30	Depressed	720
Non-depressed	77	Non-depressed	1848

Table 1: Number of speech samples in Dataset

B. Acoustic Features of Speech

Speech, back-channelling, vocal pauses, and voice quality are communicated through audio signals. These are perceived by listeners in terms of pitch, loudness, speaking rate, rhythm, voice quality, and articulation. They can be measured from recordings of spontaneous or scripted speech and quantified using a variety of parameters, such as cepstral, glottal, and spectral features. Prosodic features can be regularly identified by a listener as pitch, tone, rhythm, stress, voice quality, articulation, intonation, etc. Inspiring features in research include sentence length and rhythm, intonation, fundamental frequency, and Mel-frequency cepstral coefficients (MFCCs).

1) Audio Segmentation

Audio segmentation focuses on splitting an uninterrupted audio signal into segments of homogeneous content. An aggregated histogram of time distances between successive feature local maxima. The maximum position of the histogram is used to estimate the BPM rate. The aim of audio segmentation is algorithmic solutions for two general subcategories of audio segmentation:

- The first types of segmentation contain algorithms that utilize preceding experience, e.g. a pre-trained classification scheme. For that type of segmentation requires a fix-sized joint segmentation-classification approach and used an HMM-based method.

• The second type of segmentation is either unsupervised or semi-supervised. In both cases, no prior knowledge of the involved classes of audio content is used. Typical examples of these types of segmentation are silence removal, speaker diarization, and audio thumbnailing.

2) Feature Extraction

Details present in speech sound i.e. Acoustic features that are related to human speech production mechanism leads to classification speech into a controlled and depressed one. Acoustic features that described accurate measurements in human speech production were extracted.

Low Level Descriptors	Statistical features
normalized F0, NAQ, QOQ, H1H2, PSP, MDQ, peak Slope, Rd, Rd conf, MCEP 0-24, HMPDM 1-24, HMPDD 1-12	mean, min, skewness, kurtosis, standard deviation, median, peak-magnitude to root-mean-square ratio, root mean square level, interquartile range

Table 2: Statistical descriptors calculated from the pre-extracted audio features.

There are several ways to approach acoustic feature extraction, which is the most critical component in building a successful approach. One way of feature extraction includes extracting short-term and mid-term audio features such as MFCCs, chroma vectors, zero-crossing rate, etc. and serving them as inputs to a Classification algorithm. Since these features are lower-level representations of audio, the concern arises that complex speech features displayed by depressed individuals would go undetected.

Running a classification algorithm on the 34 short-term features yielded an encouraging F1 score of 0.59, with minimal tuning. This approach has been previously employed by others, so I treated this as "baseline" comparative data for which to develop and evaluate a completely new approach involving convolutional neural networks (CNNs) with spectrograms, which could be quite promising and powerful.

In this effort, speech stimuli are represented via a spectrogram. CNNs require a visual image. In this practice, speech motives are interpreted via a spectrogram.

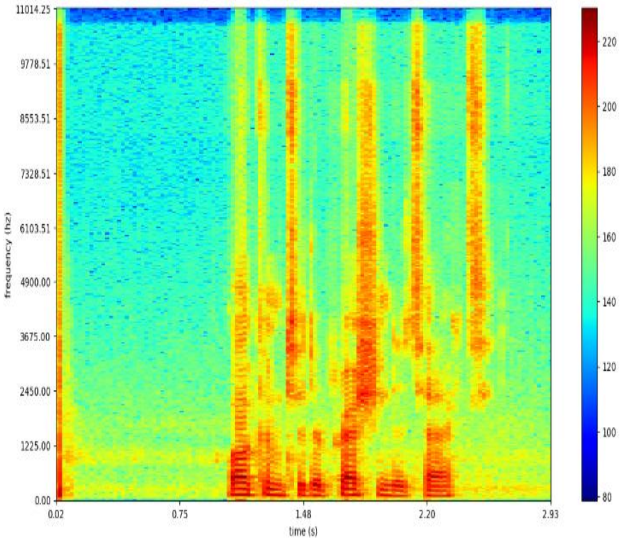

Fig -2: Spectrogram of a plosive, followed by a second of silence, and the spoken words.

Unlike MFCCs and other transformations that represent lower level features of sound, spectrograms maintain a high level of detail (including the noise, which can present challenges to neural network learning).

C. Convolutional Neural Network

CNNs have proven to be a powerful tool in image recognition, video analysis, and natural language processing. More relevant to recent work, successful applications have also been applied to speech analysis. CNNs take images as input. In the case of the spectrogram, the input of a grayscale representation, with the "grayness" a representative of the audio power level at that specific frequency and time. A filter (kernel) is consequently slid over the spectrogram image and patterns for depressed and non-depressed individuals are determined.

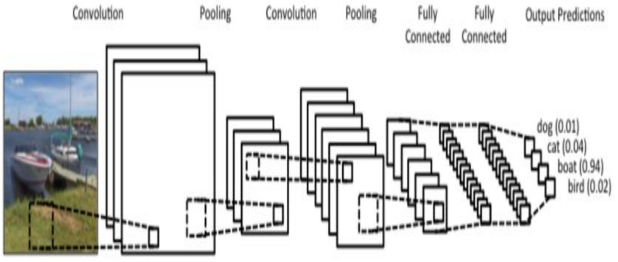

Fig -3: General CNN architecture.

Generally, a typical CNN contains one or more pairs of convolution and max-pooling layers. A convolution layer's parameters consist of a set of learnable filters. A max-pooling layer partitions convolution layer activations into non-overlapping rectangles and takes the maximum filter activation from these sub-regions. Two significant ideas are making the convolution layer useful and effective, i.e. local connectivity and weight sharing. Local connectivity restricts that each neuron connects to only a local region of inputs, leading to sparse interactions; and weight sharing reduces the number of the parameters and makes CNN much more efficient than regular feedforward multilayer perceptron's.

Features like frequency-time curve may provide a classic and powerful representation of different prosodic features of speech, which in turn are characteristic of underlying differences between depressed and non-depressed speech.

However, with the extremely accurate descriptions of speech provided in spectrograms, false noise signals (ambient noise, plosives, unsegmented audio from other speakers, etc.) can be inappropriately picked by the neural network.

$$o_j = f(\sum_{i=1}^{3} w_i x_{i+j-1}) \quad ...(1)$$

where W_i stands for the weight, and xi+j−1 is the ith input of node j, and f (·) is a non-linear activation function.

It is common to deploy a pooling layer after a convolution one, replacing the output of the net at a certain location with a summary statistic of the nearby outputs. The pooling layer aims to reduce the resolution of feature maps and introduce invariance to small variations in location. Max-pooling of local connectivity and weight sharing in the convolution network is one of the most popular pooling operations. Each node is calculated by taking the max value on the corresponding region. Pooling is a form of non-linear down-sampling, and it provides the translation invariance and tolerance to minor differences of positions of object parts, which is quite essential in this specific case of ADD since we care more about whether depression feature is present than where it is exactly in the audio signal. There exist several studies that directly learn acoustic deep models from raw waveforms as the manner in image recognition where raw pixels are usually the inputs to CNN. But recent research shows it is more efficient to build the models using some low-level audio descriptors. MFCCs are one of the most popular audio representations, whereas it is not suitable in this case since Discrete Cosine Transform (DCT) projects the spectral energies into a new basis, which may not maintain the locality required by CNN. Based on such observations, we exploit the Mel-scale filter bank features as the input. Compared to MFCCs, Mel-scale filter bank is much more appropriate for local filtering in this CNN configuration. The Mel-scale filter bank feature is computed by multiplying Short-Time Fourier Transform (STFT) magnitude coefficients with the corresponding filter, and it can thus be regarded as a non-linear transformation of a spectrogram. It builds 40 log-spaced filters according to the following Mel-scale:

$$Mel(freq) = 2595 \cdot \log_{10}(1 + \frac{freq}{700}) \quad ....(2)$$

The filter bank and spectrogram features of two audio slices in the DAIC-WOZ database are visualized.

D. Classification

Classification is the backbone of the analysis system which aims to produce the result based on the particular features of the entities constituting the dataset. Classification models were constructed to test whether voice features could be used to diagnose depression. We used feature selection, a method that improves a model's generalizability and avoids the curse of dimensionality. Data obtained after the process of pre-processing is inputted to the classifier algorithms. In the depression analysis model, we have different sets of inputs such as speech and according to different sets of these we use the different classifiers to properly classify them. In case of input in the form of speech, the model tries to recognize the words implied from speech and has added implementations of voice data which provides the information about pitch variations, intensity as well as the depth of the speaker which could imply a lot towards the determination of mental state as well as emotional stability.

The process of classification of an input dataset into result is the step by step process and may involve the depletion of procedures to attain maximum accuracy at different levels of analysis. The model of depression analysis cannot be built upon a single group of classifiers and contains several classifiers in different groups.

4. CONCLUSIONS

In this paper, We have introduced and would be implementing an artificially intelligent system that was proposed for depression level analysis using audio features. In summary, we found strong evidence that change in depression severity is revealed by vocal prosody. Listeners naive to depression scores differentiated symptom severity from the voices of participants. Specific prosodic features appeared to carry this information.

This hierarchical structure delivers a comprehensive audio representation by capturing the short term and middle-term temporal and spectral correlations with CNN. Evaluations are carried out on DAIC-WOZ used for the AVEC 2016 competition, and the results demonstrate the effectiveness of the proposed method. Experiments are performed on DAIC-WOZ, AVEC2013, and AVEC2014 datasets which are useful for depression analysis. We have studied various stages involved in the detection of depression using speech signal: signal preprocessing which involves preprocessing of the speech data obtained from the database. Feature Extraction of preprocessed speech using various features of speech such as Prosodic, Glottal, Cepstral, and Spectral. And accordingly, we found a combination of various audio features provide us with an accurate output in classifying the state of a person.

ACKNOWLEDGEMENT

These proposed work under the guidance of prof. Sonali N. Mhatre ma'am and supported by Bharati Vidyapeeth College of Engineering. We would like to thank the DAIC-WOZ for providing the depression dataset for this paper.

REFERENCES

[1] Depression article on https://www.nimh.nih.gov/health/topics/depression /index.shtml

[2] A. Jan, H. Meng, Y. F. B. A. Gaus and F. Zhang, "Artificial Intelligent System for Automatic Depression Level Analysis Through Visual and Vocal Expressions," in *IEEE Transactions on Cognitive and Developmental Systems*, vol. 10, no. 3, pp. 668-680, Sept. 2018.

[3] Dham, Shubham & Sharma, Anirudh & Dhall, Abhinav (2017). "Depression Scale Recognition from Audio, Visual and Text Analysis".

[4] Lang Hea and Cui Caob "Automated depression analysis using convolutional neural networks from speech"

[5] Cohn, Jeffrey & Kruez, Tomas & Matthews, Iain & Yang, Ying & Nguyen, Minh & Padilla, Margara & Zhou, Feng & De la Torre, Fernando. (2009). "Detecting depression from facial actions and vocal prosody. Affective Computing and Intelligent Interaction." 1 - 7. 10.1109/ACII.2009.5349358.

[6] L.-S. Low, M. Maddage, M. Lech, L. Sheeber, N. Allen, "Influence of acoustic lowlevel descriptors in the detection of clinical depression in adolescents, in: 2010 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)", IEEE, 2010, pp. 5154–5157.

[7] L.G. Hafemann, L.S. Oliveira, P. Cavalin, "Forest species recognition using deep convolutional neural networks, in: 2014 22nd International Conference on Pattern Recognition (ICPR)", IEEE, 2014, pp. 1103–1107.

[8] Yang, Ying & Fairbairn, Catharine & Cohn, Jeffrey. (2012). "Detecting Depression Severity from Vocal Prosody." IEEE Transactions on Affective Computing. 99. 10.1109/T-AFFC.2012.38.

[9] Cohn, Jeffrey & Kruez, Tomas & Matthews, Iain & Yang, Ying & Nguyen, Minh & Padilla, Margara & Zhou, Feng & De la Torre, Fernando. (2009). "Detecting depression from facial actions and vocal prosody." Affective Computing and Intelligent Interaction. 1 - 7. 10.1109/ACII.2009.5349358.

[10] Melo, Wheidima & Granger, Eric & Hadid, Abdenour. (2019). "Depression Detection Based on Deep Distribution Learning." 10.1109/ICIP.2019.8803467.

[11] Pampouchidou, Anastasia & Simos, Panagiotis & Marias, Kostas & Meriaudeau, Fabrice & Yang, Fan & Pediaditis, Matthew & Tsiknakis, Manolis. (2019). "Automatic Assessment of Depression Based on Visual Cues: A Systematic Review. IEEE Transactions on Affective Computing." 10. 445-470. 10.1109/TAFFC.2017.2724035. I. S. Jacobs and C. P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[12] Meng, Hongying & Huang, di & Wang, Heng & Yang, Hongyu & AI-Shuraifi, Mohammed. (2013). "Depression recognition based on dynamic facial and vocal expression features using partial least square regression. AVEC 2013 - Proceedings of the 3rd ACM International Workshop on Audio/Visual Emotion Challenge." 21-30. 10.1145/2512530.2512532.

[13] R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press. Jain, Varun & Crowley, James & Dey, Anind & Lux, Augustin. (2014). "Depression Estimation Using Audiovisual Features and Fisher Vector Encoding. AVEC 2014 - Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge, Workshop of MM 2014." 10.1145/2661806.2661817.

[14] Cannizzaro, Michael & Harel, Brian & Reilly, Nicole & Chappell, Phillip & Snyder, Peter. (2004). "Voice acoustical measurement of the severity of major depression. Brain and cognition." 56. 30-5. 10.1016/j.bandc.2004.05.003.

[15] Alpert, Murray & Pouget, Enrique & Silva, Raul. (2001). "Reflections of depression in acoustic measures of the patient's speech." Journal of affective disorders. 66. 59-69. 10.1016/S0165-0327(00)00335-9.

[16] Kuny, St & Stassen, Hans. (1993). "Speaking behavior and voice sound characteristics in depressive patients during recovery." Journal of psychiatric research. 27. 289-307. 10.1016/0022-3956(93)90040-9.