EasyChair Preprint

№ 8042

Speech Emotion Analysis Using Machine Learning for Depression Recognition: a Review

S Bhavya, Royson Clausit Dmello, Ashish Nayak and Sakshi S Bangera

May 22, 2022

SPEECH EMOTION ANALYSIS USING MACHINE LEARNING FOR DEPRESSION RECOGNITION: A Review

Bhavya S
Department of Electronics & Communication & Engineering
Mangalore Institute of Technology & Engineering.
Moodbidri, India
bhavyas@mite.ac.in

Royson Clausit Dmello
Department of Electronics & Communication & Engineering
Mangalore Institute of Technology & Engineering.
Moodbidri, India
roysondmello609@gmail.com

Ashish Nayak
Department of Electronics & Communication & Engineering
Mangalore Institute of Technology & Engineering.
Moodbidri, India
ashishnayak16306@gmail.com

Sakshi S Bangera
Department of Electronics & Communication & Engineering
Mangalore Institute of Technology & Engineering.
Moodbidri, India
sakshibangera44@gmail.com

***Abstract*—Depression is a psychiatric disorder which can affect individual's physical health and wellbeing. Untreated depression can disrupt a person's quality of life and results in a cascade of further symptoms. Current diagnostic approaches are confined to clinical intervention. Hence, this system is proposed to detect depression at an early stage and also to offer help while taking clinical management decisions during treatment.**

Communication is essential for conveying our thoughts and ideas to others. Machine Learning is quickly progressing in its ability to bring more sophisticated systems into everyday use. Intelligent systems are interactive and operate with little user effort, relying primarily on voice input. The purpose of this article is to show various algorithms for detecting speech emotions in order to recognize depression using machine learning.

I. Introduction

One of the most common mood disorder is depression. It is a worldwide problem affecting many lives. Depressed people find it hard to concentrate in their work and communicate with people. They may suffer from insomnia, restlessness, loss of appetite and may have suicidal thoughts. Hence depression detection is a major issue. Early intervention is very important for reducing the effects of how a person feels, thinks and behaves may lead to various emotional and physical problem. Hence a computer-based detection system is proposed as an assistance to the traditional procedures.

Communication is the key to express oneself. Speech not only conveys words and meanings but also emotions. Speech is a promising modality for identifying the human emotions. Because emotions helps to understand each other better, an algorithm to estimate depression status from a person's voice input has been developed. The emergence of Machine learning can significantly improve the diagnosis of depression. The neural networks learn from the inputs and use them to improve classification accuracy.

II. Literature Review

The main purpose of the paper is to detect depression using speech. Acoustic features are utilized to train a classification algorithm to determine whether or not a person is depressed in this article. The classifiers are trained using the DAIC-WOZ database. Prosodic, Spectral, and Voice control features are extracted using the COVAREP toolkit. The synthetic minority oversampling technique is used to eliminate the class imbalance generated by the datasets employed. Results from classification techniques such as Logistic Regression, Random Forest, and SVM are compared. CureD is an Android application for self-evaluation. By using the application a psychiatrist can perform the analysis in remote location where his physical appearance is not possible. Under the guidance of a professional psychiatrist, the application is evaluated and 90 percent accuracy is achieved. After SMOTE analysis, the SVM had the best accuracy of all the classifiers used. The model is trained to the dataset that is based on foreign accent and dataset could be extended by adding audio samples of various native accents and languages to the existing dataset [1].

A machine learning approach is used in paper [2] to suggest an automatic system. The SVM classifier is utilized in this approach to discern between healthy and sad states by examining voice data. Speech signals from psychiatrists were collected for training purposes, totaling 149 samples, 62 of which are healthy and 87 of which are problematic. All speech signals were recorded at a sample rate of 44.1 kHz with a resolution of 32 bits. Jitter, MFCC, Derivatives of cepstral coefficients, and spectral centroid are some of the acoustics properties covered. Polynomial and radial basis functions are the most commonly utilized kernel forms (RBF).Performance of this approach is evaluated on the basis of sensitivity, accuracy, precision and ROC area. The proposed simplified approach's key advantage is that it may be simply implemented in the clinical setting, even by

individuals without advanced computer abilities. This technique has an overall accuracy of roughly 85 percent.

A deep Recurrent neural network-based framework is proposed in Paper [3] to detect depression and estimate its severity level from speech. The effectiveness of this method is tested in multimodal and multifeatured tests. This system makes use of the DAIC-WOZ dataset. The recordings are split into 2 groups in this method. The audio segment of the interviewer and audio segment of participants. Audio segment of interviewer is discarded and only participants audio segment is used. Low-level features are extracted from preprocessed audio recordings and are defined as MFCC coefficients. The spoken signal is initially separated into frames using a 2.5s windowing method with 500ms intervals. The suggested model's baseline is based on LSTM. Different augmentation techniques such as noise injection, pitch augmentation, shift augmentation, and speed augmentation are used to create new audio segments. The overall accuracy of this model is 76.27%.

Based on Dempster-Shafer evidence theory, paper [4] proposes a multimodal fusion emotion recognition system. To implement this model author selected a dataset 25 video clips which consists of five types of emotions like happy, sad, relaxed, angry and disgusted. The author proposes using the DS theory to combine an ECG-based emotion recognition model with an EEG-based emotion recognition model for decision level fusion. HR and HRV features are retrieved from ECG data and categorized using a Bi-LSTM network. P-wave, Q-wave, R-wave, S-wave, and T-wave statistical properties, including mean, median, Sd, maximum, minimum, and difference between maximum and minimum values. Four types of features are utilized as input to the SVM classifier for classification of EEG signals. Based on the timing information from the previous moment, the LSTM network predicts the output of the next moment. The experimental results show that for more emotional information the fusion of EEG and ECG signal information should be used and the accuracy of emotion recognition can be improved by multimodal fusion.

Convolutional neural networks and long short term memory approaches were described in paper [5]. The DAIC-WOZ dataset is split into three parts: 80 percent for training, 10% for validation, and 10% for testing. For binary depression classes, LSTM-based audio features perform marginally better than CNN, with accuracy of 66.25 percent and 65.6 percent, respectively. In this paper, audio, visual, and textual data are explored, and they are used in multiple deep learning approaches for unimodal and multimodal representations. Only the patient's reaction is used in audio preprocessing. Activation function, recurrent step, recurrent dropout, and optimization are all employed in the LSTM layer. This model's results reveal that the proposed LSTM-based architecture outperforms the CNN-based architecture in terms of learning temporal dynamic representation of multimodal data.

independent source settings in parallel using the data given, and comparing the performance of the two alternative voice conversion methods. This study's depression analysis technique is based on thei- vector. It allows for the compression of all low-dimensional features of a voice recording, such as gender, age message, and so on. For speech synthesis and voice conversion, a Generative Adversarial Network (GAN) is used. It combines two neural networks that are in competition: a discriminator and a generator. Speaker de-identification alters a source's vocal characteristics to make it sound like a different person.

The paper [7] investigates various methods for predicting depression. Decision trees, SVM, Logistic Regression, and KNN classifiers are among examples. Wearable devices and mobile phones are used to collect behavioral data. Twint is a tool that the author uses to determine whether or not a person is depressed. Following the database extraction, a model is created and trained to achieve the desired outcome. Finally, it detects the disease via the social media platform Twitter and determines whether or not the individual is depressed. The keywords are retrieved from Twitter using the application Twint to detect depression. In this study, the author offers the Deep Learning Mechanism for detecting depression. This method, according to the author, has the potential to improve accuracy. The study also includes an example in which the author observed 46 people and identified 85 distinct traits.

The speech emotion recognition system is defined by Paper [8] as a collection of approaches for classifying speech and detecting emotions from it. The Acted Speech Emotion Database, Elicited Speech Emotion Database, and Natural Speech Emotion Database are the three elements of the database for speech emotion recognition. Preprocessing, Framing, Windowing, and Voice Activity Detection are the four processes in speech emotion recognition. After extracting the database, the first stage will be preprocessing, which will be utilized to train the model in a SER system. Speech is continually segmented into fixed length segments during signal framing, which is also known as speech segmentation. The window function is applied to the frames in the following phase. This procedure is mostly utilized during fast Fourier transforms (FFT) to minimize the effect of leakage. So, in the final stage, the created voice speech will be generated using vocal speech vibrations, which will be detected. The speech is normalized in the following steps, and any noise is recognized. Feature selection and dimension reduction are critical steps in speech emotion recognition (SER).

A real-time emotion identification system that recognizes live speech is examined in paper [9]. RAVDESS and SAVEE databases were used in this study. Basically, this work extracts and analyses 34 audio characteristics. Gradient Boosting is used to classify emotions using the trained models. Other classifiers include Support Vector Machine (SVM) and K-Nearest Neighbor (KNN). Four emotions are

Paper [6] recommends that de-identification be utilized to protect patients' privacy. This research focuses on a comprehensive investigation of depression detection utilizing voice conversion and other de-identification techniques. In this research, various aspects are examined in order to recognize depression. A few features include assessing how this system performs when the gender is changed, comparing the performance of speaker dependent and speaker examined in this book. Anger, sadness, neutrality, and happiness are all emotions. There are three key steps in this system. Features extraction, feature normalization, and categorization are examples. pyAudio Analysis extracts 34 audio attributes from each audio file. The feature normalization stage is critical because the accuracy outcome is dependent on it. Normalization aids in the weighting or normalization of all features. Gradient boosting, SVM, and

KNN are the three primary features of the third stage. In this stage, the emotions are classified using the three classifiers. According to the author, the next step in improving accuracy is to incorporate deep learning into the system.

A study on long short term memory for the diagnosis of mood disorders is presented in paper [10], which is based primarily on evoked speech replies. The MHMC emotion database was used for both evaluation and training the model. The algorithm that was used to adapt the database id HSC. The CNN technique is utilized to train the model using the altered MHMC sequences. A toolbox called openSMILE is used to extract datasets, mostly for low-level feature extraction. The emotion profile evolutions were characterized using LSTM (LP). LSTMs were created to solve the problem of long-term data storage and reliance. Because the response may contain some irrelevant information in this study, an addition method called an attention mechanism is utilized to highlight the crucial or only required content in the response.

The major goal of the paper [11] is to assess articulatory traits and factors linked with linguistic stress in clinically depressed and non-depressed speakers for automatic analysis. Brno Phoneme Recognizer is used by the author to automatically separate phonemes from audio samples. Open-source openSMILE speech toolbox is used to extract the 88 acoustic speech functional values for the acoustic feature. Both the AViD and DIAC-WOZ databases showed similar depression categorization performance patterns for tongue height, articulatory characteristics features, and advancement in terms of mid and front vowel set .Depressed speakers have shorter vowel duration and less variance for low, back, and rounded vowel locations, according to the linguistic stress feature components in both datasets. According to the articulatory qualities given, the median length for the central vowel set was the shortest among the parameters in both the DAIC-WOZ and Avid datasets, while the back, rounded, tense, low, and diphthong sets were the longest. Depressed speakers had lower loudness components and linguistic stress duration, according to experimental data utilizing various English articulatory characteristic vowel sets.

In study [12], an algorithm was developed to distinguish between depressed and non-depressed people based on their twitter status and tweets. R studio was used to do a qualitative analysis. R is a programming language designed to improve quantifiable analysis. The proposed approach is used to evaluate the twitter dataset. Data from Twitter is extracted and shown in an excel sheet. Different scores are assigned to the various emotions. Positive, negative, and natural sentiments are assessed. A score is gained for pleasant feelings, but none is obtained for negative emotions such as disgust, wrath, and so on. The sentiment values of the identical tweets are retrieved and kept. Rows and columns separate the emotions and scores. The dataset could provide detailed information about the tweets. The study is conducted using text messages, but it is also possible to do analysis using photographs and sounds.

For successful measurement of depression severity from audio samples, paper [13] suggests a combination of hand-crafted and deep learnt characteristics. The author presents a Deep Convolutional Neural Networks-based technique (DCNN). DCNNs are designed to extract deep learnt features from spectrograms and raw speech waveforms. The texture descriptors known as median robust extended local binary patterns (MRELBP) are then manually extracted from spectrograms. The author suggests a joint fine tuning layer to merge the raw speech and spectrogram based DCNN to improve the model's performance in depression recognition. It aids in the capture of additional data within the handcrafted and deeply learnt features.The studies are carried out on the depression databases AVEC2013 and AVEC2014. When compared to alternative methods based on auditory signals, the results suggest that the strategy adopted is resilient and successful for the diagnosis of depression.

A speech-based depression detection system is described in Paper [14], which can be used as a screening tool to help depressed adolescents. The software MATLAB R2010a was used to create this system. The speech signal was first pre-processed, and the retrieved characteristics produced statistically significant results. Second, the prosodic, spectral, glottal, cepstral, and Teager energy operator (TEO) categories, as well as their combinations, were studied. For training, the DAIC-WOZ database, which is part of a bigger corpus, is used. The database included 85 sessions of associations lasting between 7 and 33 minutes. When employed alone, the TEO-based features outscored all other features and feature combinations. Glottal-based feature categories were shown to have a higher accuracy of 89.41% in the Gaussian mixture model and 41.17 percent in the Support Vector Machine classifier. The author also points out that a sad adolescent may not show visible indicators of depression. As a result, there are no set criteria for detecting depression, especially when the child is establishing new roles within the family, dealing with independence, and making academic and career decisions.

Paper [15] uses a person's speech signal to assess their depression condition. When patients vocalized three types of long vowels, speech signals were gathered. The openSMILE software was then used to extract the acoustic features. Weka software was used to select the features. Finally, a 4-fold cross-validation strategy was utilized to build an algorithm that accurately measured the severity of HAM-D score from speech signals for each long vowel /Ah/, /Eh/, and /Uh/, with accuracy of 75.5 percent, 78.7%, and 68.9%, respectively.

The purpose of paper [16] is to recognize emotions. DNN (Deep Neural Network) is employed. The Convolutional Neural Network model used in this research retrieves features from a raw signal. The raw data is divided into a 20-second sequence and used as an input after processing. To extract information from the raw signal, the kernel size is 8. The maximum pooling size is determined based on the kernel size in order to reduce the signal's frame rate. To capture the information in the data, two layers of LSTM are used. The remote collaborative and affective (RECOLA) database was employed in this study.

III. Summary and Observations

According to the examined literature, multiple machine learning algorithms were used to analyze emotions from voice signals. The extraction of relevant features that may be utilized to train the machine learning model is the first step in the process. Acoustic aspects such as prosodic, spectral, glottal, and voice control were given priority. A toolkit called open SMILE is used for feature extraction, mostly for low-level feature extraction. It was also attempted to leverage multi-modal elements such as audio-visual to increase the machine learning algorithms' capacity to recognize depression.

DAIC-WOZ [18], RECOLA, TESS, and RAVDESS, all of which are freely available for research purposes, were heavily used. Some of the studies make use of datasets that incorporate data acquired from volunteers. The optimal dataset selection has a significant impact on the final result. As can be seen, dataset biases such as gender bias or age bias might have an impact on the effectiveness of the final machine learning model trained on such datasets to analyze emotions or depression. It's also crucial to keep a healthy ratio between sad and nondepressed data, as well as across samples related to other emotions, so that the model isn't distorted by majority sample bias.

It should also be mentioned that appropriate models for predicting emotional information from speech must be created using suitably large emotional speech corpuses. The creation of accurate prediction models and the construction of an adequate emotional speech corpus are the main concerns. The scarcity of data for the study makes it challenging to develop improved models. It's also tough to share information on depression recognition because it's so private.

The investigation employed machine learning algorithms such as SVM, Decision Trees, Logistic Regression, and KNN classifiers. The SMOTE approach was employed to improve the model's accuracy. Deep learning models such as CNN, ANN, and LSTM, on the other hand, outperformed traditional machine learning approaches.

Paper	Dataset	Algorithms	%Accuracy
[1]	DAIC-WOZ	SVM	70.2
		Random Forest	59.7
		Logistic Regression	63.8
[3]	DAIC-WOZ	LSTM	76.27
[5]	DAIC-WOZ	MFCC-AU LSTM	95.38
[14]	DAIC-WOZ	SVM	70.58
		Gaussian Mixture Model	88.23

Table.1 Performance of different methods on dataset

The models mentioned were trained on a dataset based on foreign accents, however the dataset might be expanded by adding audio samples of other native accents and languages to the existing datasets, extending the research's scope beyond linguistic barriers.

IV. Conclusion

Depression is a significant medical condition. The proposed system is designed to reduce human intervention in the process of diagnosing depression in an individual. Speech is regarded as essential. The lack of publicly available speech databases made developing a well-trained model difficult. Keras and Scikit Learn are used to create the depression recognition system. The Distress Analysis Interview Corpus (DAIC) database was used to detect clinical depression in speech.

To increase system performance, future developments in the proposed system would include implementing the model utilizing a large audio dataset, as well as merging face expression information recorded via video and adding languages for diagnoses, namely the local languages spoken in India. We also hope to integrate our proposed technology into a real-time depression detection program to provide emotional support.

References

[1] B. Yalamanchili, N. S. Kota, M. S. Abbaraju, V. S. S. Nadella and S. V. Alluri, "Real-time Acoustic based Depression Detection using Machine Learning Techniques," 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE), 2020, pp. 1-6.

[2] L. Verde et al., "A Lightweight Machine Learning Approach to Detect Depression from Speech Analysis," 2021 IEEE 33rd International Conference on Tools with Artificial Intelligence (ICTAI), 2021, pp. 330-335, doi: 10.1109/ICTAI52525.2021.00054.

[3] Emna Rejaibi et al., "MFCC-based Recurrent Neural Network for automatic clinical depression recognition and assessment from speech", Biomedical Signal Processing and Control, Volume 71, Part A, 2021, 103107, ISSN 1746-8094, doi.org/10.1016/j.bspc.2021.103107.

[4] Tian Chen, Hongfang Yin , Xiaohui Yuan , "Emotion recognition based on fusion of long short-term memory network and SVMs" , Digital signal processing , School of Computer Science and Information Engineering 230009 (2021).

[5] Muhammad Muzammel, Hanan Salam, Alice Othmani, "End -to -end multimodal clinical depression recognition using deep neural networks: A comparative analysis.", Computer Methods and Programs in Biomedicine (UPEC), LISSI, Vitry Sur Seine 94400 -2021

[6] Paula Lopez-Otero, Laura Docio-Fernandez, "Analysis of gender and identity issues in depression detection on de-identification speech", Computer Speech & Language, E.E Telecommunication campus – 2021 dio.org/10.1016/j.csl.2020.101118.

[7] P. V. Narayanrao and P. Lalitha Surya Kumari, "Analysis of Machine Learning Algorithms for Predicting Depression," 2020 International Conference on Computer Science, Engineering and Applications (ICCSEA), 2020, pp. 1-4.

[8] Mehmet Berkehan Akcay, Kaya Oguz , " Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers" , Speech Communication, Department of Computer Science Engineering doi.org10.1016/j.specom.2019.12.001 (2020).

[9] Iqbal, A. and Barua, K. "A real-time emotion recognition from speech using gradient boosting" In 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE) (2019), IEEE, pp. 1–5

[10] Kun-Yi Huang Chung-Hsien Wu, Ming -Hsiang Su, " Attention based convolutional neural network and long short-term memory for short -term detection of mood disorders based on elicited speech responses." ,Pattern recognition , Department of Computer Science and Information Science Engineering. 0031-3203(2019).

[11] Brain Stasak , Julien Epps, Roland Goecke , " An investigation of linguistic stress and articulatory vowel characteristics for automatic depression classification" , Computer speech and Language 53(2019) 140-155 .

[12] A. Sood, M. Hooda, S. Dhir and M. Bhatia., "An Initiative to Identify Depression using Sentiment Analysis: A Machine Learning Approach," Indian Journal of Science and Technology, 2018, Vol 11(4).

[13] Lang He, Cui Cao., "Automated depression analysis using convolutional neural networks from speech" , Journal of Biomedical Informatics , NPU-VUB joint AVSP Research Labs, dio.org/10.1016/j.jbi.2018.05.007 -2018.

[14] P. R. Parekh and M. M. Patil, "Clinical depression detection for adolescent by speech features," 2017 International Conference on Energy, Communication,

Data Analytics and Soft Computing (ICECDS), 2017, pp. 3453-3457.

[15] Y. Omiya et al., "Estimating depressive status from voice," 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2018, pp. 2795-2796

[16] Panagiotis Tzirakis ,Jiehao Zhang , Bjorn W. Schuller, " End-to-end Speech Emotion Recognition using Deep Neural Networks". Department of computing Imperial College, 978-1-5386-4658-8, 2018 IEEE

[17] Livingstone SR, Russo FA (2018) The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. PLoS ONE 13(5): e0196391.2018, https://doi.org/10.1371/journal.pone.0196391,201

[18] Jonathan Gratch, Ron Artstein, Gale Lucas, Giota Stratou, Stefan Scherer, Angela Nazarian, Rachel Wood, Jill Boberg, David DeVault, Stacy Marsella, David Traum, Skip Rizzo, Louis-Philippe Morency, "The Distress Analysis Interview Corpus of human and computer interviews", Proceedings of Language Resources and Evaluation Conference (LREC), 2014