A Novel Technique for Detecting Depressive Disorder: A Speech Database-Based Approach

Bubai Maji[1], Anup Kumar Roy[2], Shazia Nasreen[3], Rajlakshmi Guha[3], Aurobinda Routray[4], and Debabrata Majumdar[5]

***Abstract*—Depression is the second most diagnosed disease in the world and is predicted to be the highest by the year 2030. Depressive disorder impacts both on mentally and physically, thus diagnosing this disorder in early stage is essential. Automatic Depression Detection (ADD) system via speech can greatly facilitate early-stage depression diagnosis. Development of such systems demands a standard balanced database. In this work, we present a novel labeled audio distress interview database. To our knowledge, this is the first depression database in Bengali language that contains audio responses from depressed and non-depressed subjects. Alongside this, we present a set of hand-crafted acoustic features that effectively detect depression mood using speech signals. Finally, we justify the quality of our developed database and the efficacy of the feature set in predicting depression using a baseline machine learning (ML) model. We believe that the annotated database will be a valuable resource for use by treating clinicians.**

***Clinical Relevance*—This research reports a new speech database in Bengali language for depression detection. This database can be used in healthcare by developing an automatic prediction model for depression detection.**

I. Introduction

Depression can affect a person's ability to engage in social activities and communication. Severe depression can be one of the main reasons for suicide [1]. The World Health Organization (WHO) estimates that by 2030, depression disorder will become the world's first disease burden [2]. This can be prevented if depressed people go for help from health professionals and if automated tools could assist in screening and diagnosis. Currently, clinicians diagnose depression by clinical interviews and using different psychological scales like, the Hamilton depression rating scale (HAM-D) [3] and the Beck depression inventory scale (BDI) [4]. Automated tools on depression focus on using chatbots, textual [5], visual information [6], and EEG signals [7] but use of speech signals for depression detection is rare [8]. Among those approaches, automatic depression detection (ADD) using speech signals has attracted more attention because speech signals can be an effective clinical marker for depression [1], [9]. While in the past decade, Automatic Mood Disorder Detection technique has been widely used, research in Automatic depression detection using speech features is comparatively lesser [8]. Early studies reported that the speech characteristics, such as lower speaking rate and repeated pauses, compared to non-depressed speech [1] [5] [10]. Thus, these markers in speech can provide an alternative pathway to differentiate the depression level in speech. So, extracting useful depression-related features from the speech signal is one of the vital challenges in developing an ADD system.

Most of the recent studies use hand-crafted features such as formants [10], prosody [11], Mel Frequency Cepstral Coefficients (MFCCs) [12], Mel-spectrogram [9], voice quality [13], and vocal tract coordination (VTC) [14] to detect depression severity level from speech cues. Our objective is to analyze and demonstrate different acoustic features that give the best results for recognizing depression. This may help build an effective ADD system that supports clinicians in diagnosing depression. However, building such systems requires a balanced, high quality large datasets. Currently, several depression databases have been released [15]-[22], as shown in Table 1.

With the rapid development of speech-based computational models, in 2013, the Audio-Visual Emotion Recognition Challenge (AVEC2013) [18] and AVEC2014 [19] collected depression data using the interaction between human-computer. This dataset comprises audio and video recordings of German volunteers answering questions separately divided into training, validation, and test portions. Another popular dataset is the Distress Analysis Interview Corpus (DAIC) [20]. It consists of recordings and transcripts of American subjects which a computer agent clinically interviewed to assist the diagnosis of mental disorders such as depression, anxiety, and post-traumatic stress. Recently in 2022, Emotional Audio-Textual Depression (EATD) dataset [22] was developed in the Chinese language. It contains the recording of audio and transcripts extracted from the subjects' interviews.

However, the depression databases used in previous works suffer from two key limitations. Firstly, the unbalanced size of sample in the existing databases which may degrade the performance of these models. For example, the size of the data in the EATD corpus has only 30 depression volunteers and 132 non-depressive volunteers, which is far from the balance number. Secondly, most datasets are built on limited preset question per subject, which may fail to perform in real-time scenario. Thus, development of such a dataset without preset questions remains a challenging task.

[1,3]Rekhi Centre of Excellence for the Science of Happiness, Indian Institute of Technology Kharagpur, India (e-mail: bubaim@kgpian.iitkgp.ac.in, shaziacmc@gmail.com, rajg@cet.iitkgp.ac.in)

[4]Department of Electrical Engineering, Indian Institute of Technology Kharagpur, India (e-mail: aurobinda.routray@gmail.com).

[2]School of Medical Science and Technology, Indian Institute of Technology, Kharagpur, India (e-mail: anuproy00029@gmail.com).

[5]Consultant Psychiatrist B.C. Roy Tech Hospital Counseling Center, Indian Institute of Technology Kharagpur, India (e-mail: mdebabrata@gmail.com).

TABLE I. COMPARISON TO EXISTING PUBLIC DEPRESSION DATASETS.DPRD-DEPRESSED, A-AUDIO, V-VISUAL, T-TEXUAL,DSM-DIAGNOSTIC AND STATISTICAL MANUAL OF MENTAL DISORDERS, PHQ-9 -PATIENT HEALTH QUESTIONNAIRE,HAMD-HAMILTON RATING SCALE FOR DEPRESSION,BDI-BECK DEPRESSION INVENTORY

Database	Published Year	Language	Subjects	Modality	Annotation	Ground truth
[15]	2004	American English	33	A	DSM-IV	Clinical assessment
[16]	2009	American English	57	A+V	DSM-IV, HAMD ≥ 15	—
[17]	2012	English	60	A	DSM-IV, HAMD >15	Clinical assessment
[18]	2013	German	292	A+V	BDI-II	Self-report
[19]	2014	German	292	A+V	BDI-II	Self-report
[20]	2014	English	142	A+V+T	DPRD = PHQ-9 > 10	Self-report
[21]	2018	African-American	57	A+V	DSM-IV, HAMD >15	Clinical assessment
[22]	2022	Chinese	162	A+T	SDS × 1.25 ≥ 53	Self-report
Our database	2023	Indic-Bengali	58	A	DSM-V	Clinical assessment

To overcome these challenges, we introduce a novel dataset that is rich in varied unrestricted free-flowing content and does not limit the responses of participants to answering direct questions. The dataset comprises audio responses from 58 subjects' interviews consisting of depressed and non-depressed participants. Besides, we present different types of acoustic features extracted from the dataset. Finally, we benchmark our database with a ML model that uses extracted features; to illustrate the dataset is well-suited for ML-based methods, which would help psychologists to diagnose patients' depression.

The rest of the paper is arranged as follows. Section 2 presents a details description of our dataset. The details implementations of the baseline experiments are discussed in Section 3. Results and discussion are presented in Section 4, and Section 5 concludes the paper.

II. DATASET

To date, very few datasets refer to depression detection, as shown in Table 1. In this work, we created a new speech-based depression database in the Indic Bengali language to assist the researcher in depression detection. We planned to make the database freely available to the research community upon request.

A. Ethical Approval

We have collected ethical approval (No. IIT/SRIC/DEAN/2023) for our experimental design from the Institute's ethical committee.

B. Participants

Ongoing depressed subjects were conducted at Dr. B. C. Roy Hospital, IIT Kharagpur, and normal subjects were college students at IIT Kharagpur, India. All subjects were asked to sign an informed consent and be aware of the study's objectives. In addition, the participants' selection was made by excluding those suffering from a psychiatric disorder, severe somatic diseases, alcohol, pregnant woman, and drug abusers. As of now, our dataset consists of 23 depressed subjects and 35 normal subjects with an age range of 21-32. Furthermore, participants were examined by a clinical psychiatrist following the Diagnostic and Statistical Manual of Mental Disorders (DSM-V) [23]. The details distribution of the participants is shown in Table 2.

C. Materials

Each Participant was seated in front of a computer screen where different unstructured Rorschach Inkblot cards (RIBT) were displayed [24] [25]. Participants were asked to respond to the card by saying what they understood from the pictures. Initial and inquiry phases were completed according to Klopfer's data collection protocol. Participants' response was recorded using Sony ICD-UX570F Light Weight Voice Recorder at a sampling rate of 44.1 kHz with a 16-bit quantization rate and saved in WAV format. The average duration of the experiment is 40±4.6 minutes. The whole experiment was performed in a clean and acoustic room. All the data collection procedure and the recorded responses were validated by a clinical psychologist.

TABLE II. DATA DISTRIBUTION

	Normal	Depression	Total
Male	25	12	37
Female	10	11	21
Total	35	23	58

D. Data Preparation

Various preprocessing steps were performed on the recorded audio files. First, silent audios, mute segments at the starting and the end of every recorded file, and audio response less than 1 second are removed. Then each file is segmented into 7- 25 seconds with a total segment of 1179.

III. BASELINE EXPERIMENTS

A. Input Features

This section introduces a set of hand-crafted acoustic features extracted from our developed depression database. The extracted different features found in previous psychiatric literature to detect the depression effectively from speech signals [17] [27]. The feature set incorporate the following:

Spectral features: 13-dimensional MFCCs, 40-dimensional Mel-spectrograms, spectral centroid (SC), and zero-crossing rate (ZCR) extracted using Librosa library [28] with 25msec window lengths and a 10 msec frame interval.

Voice-related features: Fundamental frequency (F0), Formants (F1, F2, F3), Harmonic-to-Noise Ratio (HNR),

shimmers, and jitters. Praat toolkit [29] has been used to extract these features.

B. Feature Standardization

After the extracted of acoustic features from the raw data, we standardized all features using z-scores, i.e., subtracting mean and dividing by the standard deviation. The standard z-score of samples x is computed as:

$$z = \frac{x-\mu}{\sigma} \quad (1)$$

Here μ and σ are defines the mean and standard deviation of all training samples.

C. Model Training

The speech was classified in a binary event (i.e., depressed/normal). The Support Vector Machine (SVM) is well-suited and widely used of capturing the temporal information from the acoustic features. Following the approach of many speech-based classification tasks, the SVM classifier was used [17] [22]. To show the efficacy of our database in predicting depression, we also use an SVM classifier with a linear kernel, separately on the set of extracted acoustic features. Model hyper-parameters are set to their default values for all experiments as on the Scikit-learn implementation [30]. Experiments are repeated 5 times with randomly initialized weights. We split the data into 5-fold; at each fold, we have a train, validation, and test set in the ratio of 8:1:1.

D. Evaluation Metrics

The performance of the baseline ML model was analyzed using the following four metrics: (1) Accuracy; (2) Recall: the ratio of actual predicted positive samples to all positive samples; (3) Precision: the ratio of accurately predicted positive samples to the total predicted positive samples; and (4) F1 score: the harmonic mean of Recall and Precision.

IV. RESULTS AND DISCUSSION

We report and discuss the results of the baseline model (SVM classifier) of various input features, i.e., spectral and voice-related features, using our depression speech data. Table 3 represents the precision, recall, F1-score, and accuracy for each acoustic feature and combined features analyzed by the SVM classifier.

We noticed that the recognition results for the 13-MFCCsand all combined features were very competitive and significantly higher than the other features. The use of only MFCCs could also reduce the model training instead of multiple acoustic features. It suggests that the MFCC features are more effective for depression recognition than combining multiple features from speech. This result is consistent with [17], which analyzed data from the Black Dog Institute dataset; here, our investigation is only on the speech sample. The Mel-spectrogram and SC features were relatively good for depression classification, while shimmer, jitter, F0, and HNR were the worst at detecting depression.

It is noteworthy that the speech samples did not undergo any preprocessing or data enhancement process before extracting the features. Yet, the model trained on acoustic features is competitive; indicating recordings quality of the data is sufficient and is an important base for future research work. This justifies the robustness of the spectral and voice-related features in depression recognition task and the quality of the developed database. However, these speech features may be promising for depression recognition in other languages.

TABLE III. EXPERIMENTAL RESULTS OF DIFFERNET ACOUSTIC FEATURES ON OUR DATABASE

Feature	Precision	Recall	F1-Score	Accuracy
13-MFCCs	0.668	0.801	0.729	0.728
40-MFCCs	0.647	0.792	0.712	0.704
Mel-spectrogram	0.829	0.318	0.459	0.659
SC	0.742	0.667	0.704	0.710
ZCR	0.566	0.645	0.603	0.612
F0	0.545	0.442	0.488	0.605
Formants	0.563	0.531	0.547	0.623
Jitter	0.765	0.137	0.232	0.614
Shimmer	0.607	0.179	0.276	0.600
HNR	0.614	0.284	0.388	0.618
Average	0.654	0.479	0.508	0.647
Combined	0.747	0.703	0.724	0.758

We further investigate the differences in predictive accuracy indifferent genders, as illustrated in Fig 1. We observe that recognition of depression in female subjects was better, with an accuracy of 83.2%, while males had an accuracy of only 76.3%, with a mixed gender result of 72.8%. This result confirms previous conclusions of gender differences in depression detection [17]. This may be because women are more likely to vocalize their moods, as compared to men.

However, it is also important to discuss the limitations of the current study. First, creating a database and annotation process is time-consuming as many patients of depression are unwilling to allow recording of their speech content. It is also important that the audio recording itself does not act as a variable to determine the sharing of content. Secondly, the external validation with other independent datasets with different classifiers was not analyzed in this study. So, validating our findings at large and independent data with different classifiers is the next critical study.

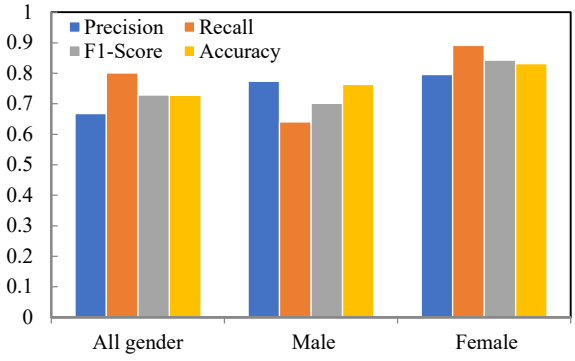

Figure 1. Performances in gender variations in terms of Accuracy, Precision, Recall, and F1-Score using MFCCs.

V. Conclusion and Future Work

We have successfully developed the first balanced database for depression detection in Bengali language and analyzed multiple acoustic features for predicting depression. We plan to enlarge the dataset with a larger and richer representative sample including other forms of depressive disorders like Dysthymia and Mixed anxiety depression. We shall also test our database with existing standardized datasets to evaluate the model's sensitivity to different languages.

References

[1] N. Cummins, S. Scherer, J. Krajewski, S. Schnieder, J. Epps, and T. F. Quatieri, "A review of depression and suicide risk assessment using speech analysis," *Speech Commun.*, vol. 71, pp. 10–49, Jul. 2015.

[2] C. D. Mathers, D. Loncar, "Projections of Global Mortality and Burden of Disease from 2002 to 2030," *PLoS Med* 3(11): e442.2006. https://doi.org/10.1371/journal.pmed.0030442

[3] M. Hamilton and W Guy, "Hamilton depression scale," *Group*, 1976, 1–4.

[4] A. T. Beck, R. A. Steer, and M. G. Carbin, "Psychometric properties of the beck depression inventory: Twenty-five years of evaluation," *Clinical psychology review*, vol. 8, no. 1, pp. 77–100, 1988.

[5] D. J. France, R. G. Shiavi, S. Silverman, and M. Silverman, "Acoustical properties of speech as indicators of depression and suicidal risk," *IEEE transactions on bio-medical engineering*, vol. 47, p. 829, 2000.

[6] A. Pampouchidou, P. Simos, K. Marias, F. Meriaudeau, F. Yang, M. Pediaditis, and M. Tsiknakis, "Automatic assessment of depression based on visual cues: A systematic review," *IEEE Trans. Affect. Comput*., p. 1, 2018.

[7] X. Li, B. Hu, S. Sun, and H. Cai, "EEG-based mild depressive detection using feature selection methods and classifiers," *Computer Methods & Programs in Biomedicine*, vol. 136, pp. 151–161, 2016.

[8] P. Lopez-Otero, L. Docio-Fernandez, and C. Garcia-Mateo, "A study of acoustic features for the classification of depressed speech," *in MIPRO. IEEE*, pp. 1331–1335, 2014.

[9] J. Wang, V. Ravi, J. Flint, and A. Alwan, "Unsupervised Instance Discriminative Learning for Depression Detection from Speech Signals," in *Proc. INTERSPEECH*, pp. 2018–2022, 2022.

[10] N. Cummins, J. Epps, M. Breakspear, and R. Goecke, "An investigation of depressed speech detection: Features and normalization," in *Proc. INTERSPEECH*, pp. 27–31, 2011.

[11] Y. Yang, C. Fairbairn, and J. F. Cohn, "Detecting depression severity from vocal prosody," *IEEE Transactions on Affective Computing*, vol. 4, no. 2, pp. 142–150, 2012.

[12] E. Rejaibi, A. Komaty, F. Meriaudeau, S. Agrebi, and A. Othmani, "MFCC-based recurrent neural network for automatic clinical depression recognition and assessment from speech," *Biomedical Signal Processing and Control*, 71, 103107, 2022.

[13] S. Scherer, G. Stratou, J. Gratch, and L. P. Morency, "Investigating voice quality as a speaker independent indicator of depression and ptsd," in *Proc. INTERSPEECH*, pp. 847–851, 2013.

[14] Z. Huang, J. Epps and D. Joachim, "Exploiting Vocal Tract Coordination Using Dilated CNNS For Depression Detection in Naturalistic Environments," *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, pp. 6549–6553, 2020.

[15] E. I. I. Moore, M. Clements, J. Peifer and L. Weisser, "Comparing objective feature statistics of speech for classifying clinical depression," *The 26th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, San Francisco, CA, USA, pp. 17–20, 2004.doi: 10.1109/IEMBS.2004.1403079.

[16] J. F. Cohn et al., "Detecting depression from facial actions and vocal prosody," *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*, Amsterdam, Netherlands, pp. 1–7, 2009. doi: 10.1109/ACII.2009.5349358.

[17] S. Alghowinem, et al., "From joyous to clinically depressed: Mood detection using spontaneous speech," in Proc. *25th Int. Florida Artif. Intell. Res. Soc. Conf*., pp. 141–146, 2012.

[18] M. Valstar, et al., "AVEC 2013: The continuous audio/visual emotion and depression recognition challenge," in Proc. *3rd ACM Int. Workshop Audio/Visual Emotion Challenge*, pp. 3–10, 2013.

[19] M. Valstar, et al., "AVEC 2014: 3D dimensional affect and depression recognition challenge," in Proc. *4th Int. Workshop Audio/Visual Emotion Challenge*, pp. 3–10, 2014.

[20] J. Gratch, R. Arstein, G. M. Lucas, G. Stratou, and S. Scherer, "The distress analysis interview corpus of human and computer interviews," in Proc. *LREC* 2014, p. 3123–3128, 2014.

[21] H. Dibeklioğlu, Z. Hammal and J. F. Cohn, "Dynamic Multimodal Measurement of Depression Severity Using Deep Autoencoding," in *IEEE Journal of Biomedical and Health Informatics*, vol. 22, no. 2, pp. 525–536, March 2018.

[22] Y. Shen, H. Yang, and L. Lin, "Automatic Depression Detection: An Emotional Audio-Textual Corpus and A Gru/Bilstm-Based Model," *ICASSP 2022 - 2022 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, pp. 6247–6251, 2022.

[23] D. Regier, W. Narrow, E. Kuhl, and D. Kupfer, "The conceptual of DSM-V," *American Journal of Psychiatry*, vol. 166, pp. 645–650, 2009.

[24] A. K. Roy, S. Nasreen, D. Majumder, M. Mahadevappa, R. Guha, and J. Mukhopadhyay, "Development of Objective Evidence in Rorschach Ink Blot Test: An Eye Tracking Study," *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Berlin, Germany, pp. 1391–1394, 2019.

[25] S. Nasreen, A. K. Roy, and R. Guha, "Exploring 'Little-c' Creativity Through Eye-parameters,"*2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, Glasgow, Scotland, United Kingdom, pp. 1078–1081, 2022.

[26] F. Ringeval et al., "Avec 2017: Real life depression, and affect recognition workshop and challenge," In Proceedings of the *7th Annual Workshop on Audio/Visual Emotion Challenge*, pp. 3–9, 2017.

[27] D. M. Low, K. H. Bentley, and S. S. Ghosh, "Automated assessment of psychiatric disorders using speech: A systematic review," *Laryngoscope Investigative Otolaryngology*, vol. 5, pp. 96–116, 2020.

[28] B. McFee et al., "librosa: Audio and music signal analysis in python," in Proc. *Python Sci. Conf*., vol. 8, pp. 18–25, 2015.

[29] P. Boersma and V. Van Heuven, "Speak and unspeak with praat," *Glot International*, vol. 5, pp. 341–347, 2001.

[30] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.