An Automatic Depression Recognition Method from Spontaneous Pronunciation Using Machine Learning

Minghao Du
Academy of Medical Engineering and Translational Medicine, Tianjin University, Tianjin, China 300072
minhodu@tju.edu.cn

Wenquan Zhang
Academy of Medical Engineering and Translational Medicine, Tianjin University, Tianjin, China 300072
zwq_@tju.edu.cn

Tao Wang
Academy of Medical Engineering and Translational Medicine, Tianjin University, Tianjin, China 300072
taowang2021@tju.edu.cn

Shuang Liu*
Academy of Medical Engineering and Translational Medicine, Tianjin University, Tianjin, China 300072
shuangliu@tju.edu.cn

Dong Ming
College of Precision Instruments & Optoelectronics Engineering, Tianjin University, Tianjin, China 300072
richardming@tju.edu.cn

ABSTRACT

The rapidly growing number of depressed people increases the burden of clinical diagnosis. Due to the abnormal speech signal of depressed patients, automatic audio-based depression recognition has the potential to become a complementary method for diagnosing. However, recognition performance varies largely with different speech acquisition tasks and classifiers, making results not comparable, and the performance requires further improvement before clinical application. This work extracted high-level statistical acoustic features (prosodic, voice-quality, and spectral features) of 23 depressed patients and 29 healthy subjects under spontaneous pronunciation tasks (interview and picture description) and mechanical pronunciation tasks (story reading and word reading), then applied principal component analysis (PCA) to reduce features dimensions, finally employed multilayer perceptron (MLP) to establish the classification model and compared with traditional classifiers (logistic regression, support vector machine, decision tree, and naive Bayes). The results showed that spontaneous pronunciation induced more significantly discriminative acoustic features and achieved better recognition performance accordingly. And the PCA retained 90% useful information with 50% features. Furthermore, MLP achieved the best performance with the accuracy 0.875 and average F1 score 0.855 under the picture description task. This study provides support for task design and classifier building for audio-based depression recognition, which could assist in mass screening for depression.

*Corresponding authors.

ICBBE 2022, November 10–13, 2022, Kyoto, Japan

ACM ISBN 978-1-4503-9722-3/22/11…$15.00
https://doi.org/10.1145/3574198.3574219

CCS CONCEPTS

• **Computing methodologies**; • **Machine learning**; • **Learning paradigms**; • **Supervised learning**; • **Supervised learning by classification**;

KEYWORDS

Depression recognition, Audio, Machine learning, Speech task

ACM Reference Format:
Minghao Du, Wenquan Zhang, Tao Wang, Shuang Liu, and Dong Ming. 2022. An Automatic Depression Recognition Method from Spontaneous Pronunciation Using Machine Learning. In *2022 9th International Conference on Biomedical and Bioinformatics Engineering (ICBBE 2022), November 10–13, 2022, Kyoto, Japan.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3574198.3574219

1 INTRODUCTION

Depression is one of the common and serious psychological disorders characterized by persistent pessimism, cognitive decline, and social dysfunction [1]. According to the World Health Organization, an estimated 3.8% of the population suffers from depression, which severely increases the burden of diagnosis [2]. In addition, existing clinical diagnoses mostly rely on subjective scales with low efficiency and high subjectivity [3]. Therefore, providing an effective, objective, and convenient method as a complement to improve current diagnostic capabilities is of vital significance.

Existing studies have indicated that depression affects cognitive function and muscular coordination [4][5], which in turn affects the physiological process of pronunciation, and patients are often characterized by speaking less, low volume, and hesitation [6][7]. Therefore, audio streams contain depression-related information that can be used for classification. More importantly, the contact-free audio collecting process is friendly to patients. Recently, several studies have explored audio-based recognition for depression. For speech collecting tasks, designing a standard task with a well-inducing effect is critical and difficult, and much effort has been made by researchers. According to the openness of questions, existing tasks can be divided into spontaneous pronunciation and mechanical pronunciation. Spontaneous pronunciation is that subjects give inconsistent answers to open questions based on their

own experience and understanding. It contains the active thinking of the subjects, but the unfixed response makes analysis difficult. For example, Gratch et al. [8] simulated the clinical interview task by a human-computer interaction system. An animated interviewer could ask several questions in a variable order and record responses. And mechanical pronunciation is that subjects read fixed materials. It standardizes the acquisition process but reduces the interaction process. For example, Valstar et al. [9] selected a short story and songs as materials to ask subjects to pronounce in their common ways. And Taguchi et al. [10] designed a speech task that subjects should read out ten digits "012–345–6789" like a telephone number and analyzed the difference of frequency features between depressed and non-depressed groups. Different effects induced by these two kinds of tasks will affect the subsequent recognition effect, which is seldom discussed in previous studies. Therefore, the selection of speech tasks is needed for ensuring the quality of depression-related expressions and reducing the interference of irrelevant emotions.

By extracting utterance-level features as global acoustic features, researchers establish several machine learning models for classification. For classifiers, Valstar et al. [11] built a support vector machine (SVM) based on the eGeMAPS features and obtained the F1 score 0.57. The eGeMAPS feature is a standardized acoustic feature set for emotional computing, but its high dimension puts forward higher requirements for the fitting ability of the classifier. Besides, Deshpande et al. [12] extracted a broad spectrum of features derived from audio samples, then proposed a random forest model to detect emotional valence and achieved accuracy 0.70. Pan et al. [13] built a logistic regression (LR) model based on multiple hand-craft acoustic features and further improved the depression recognition performance with accuracy 0.83. Instead of machine learning models, Ma et al. [14] proposed DepAudioNet framework based on deep learning and achieved F1 score 0.61, which indicated the advantages of neural networks in processing acoustic features. However, due to the difference of speech collecting tasks, the performance of these different classifiers is not comparable. Besides, because of the difficulty of fitting high-dimensional acoustic features, the performance requires further improvement before clinical application.

Considering the problems mentioned above, this work compared the induction effects of four speech tasks (interview, story reading, words reading and picture description), then proposed an automatic depression recognition method from spontaneous pronunciation using machine learning. First, we extracted acoustic features (prosodic, voice-quality and spectral features) as a low-level descriptor (LLD) from 23 depressed patients and 29 healthy subjects, and calculated globally high-level statistical features (HSF) as utterance-level features. Second, we statistically analyzed the differences in acoustic features between the depressed group and non-depressed groups under different speech tasks and applied principal component analysis (PCA) to reduce dimensions. Finally, we employed multilayer perceptron (MLP) to establish the classification model and compared it with traditional classifiers including LR, SVM, decision tree (DT) and naive Bayes (ND). Finally, we further determined the optimal combination of the speech task and classifier for the performance improvement of audio-based depression recognition.

2 DATASET

For the experiments, we used the raw audio collected from a multimodal open dataset for mental-disorder analysis (MODMA) at Lanzhou University, China [15]. This dataset also collects electroencephalogram signals for analysis [16] but is not used in this study. It contained 52 subjects including 23 depressed outpatients and 29 healthy subjects (1 depression data defective). Subjects were classified as depressed and non-depressed by clinical psychiatrists using a combination of scores from several psychological scales, including GAD-7 and PHQ-9 as shown in Figure 1. During the experiment, subjects were asked to complete the speech tasks in turn following the instructions shown on a computer screen. The spoken language was Chinese. The whole experiment lasted about 25 minutes. All audio was recorded by Neumann TLM102 (microphones) and RME FIREFACE UCX (audio card) with a 44.1 kHz sampling rate in uncompressed WAV format. Participants were asked to complete the following four speech tasks:

- **Interview.** Subjects should answer 18 questions with three kinds of valences: positive, neutral, and negative, such as "What is the best gift you have ever received and how did you feel? How do you evaluate yourself?".
- **Story reading.** The material is a short story with neutral valence named "The North Wind and the Sun", the English version of which is also used in [8] before. The subjects can read aloud in their common habits.
- **Word reading.** There are three groups of words that correspond to three kinds of valences. Each group has six words, such as "happy, center, wail".
- **Picture description.** It includes three pictures of different facial expressions and one picture from The Apperception Test (TAT), which is widely used to reflect individual psychology. Subjects are free to describe according to their own understanding.

3 METHOD

The overall process of depression recognition consisted of five parts: preprocessing, feature extraction, dimension reduction, classification and evaluation. Besides, statistical analysis is also used to analyze the discrimination of acoustic features for four speech tasks, as shown in Figure 2.

3.1 Preprocessing

Because the audio was recorded separately, all segments of the same tasks were spliced randomly to counteract the sequence effect. Then we reduced the sampling rate to 16 kHz and randomly divided all samples into the training set and test set at the ratio of 7:3. And to overcome the problem of data scarcity, data augmentation was used to increase the sample size by dividing the audio into 3 equal parts. Such augmentation was only used for training set and not for test set and statistical analyses. Finally, for statistical analysis, there were 22 subjects in the depressed group and 29 subjects in the non-depressed group, and for classification, there were 105 samples in the training set and 16 samples in the test set for each task.

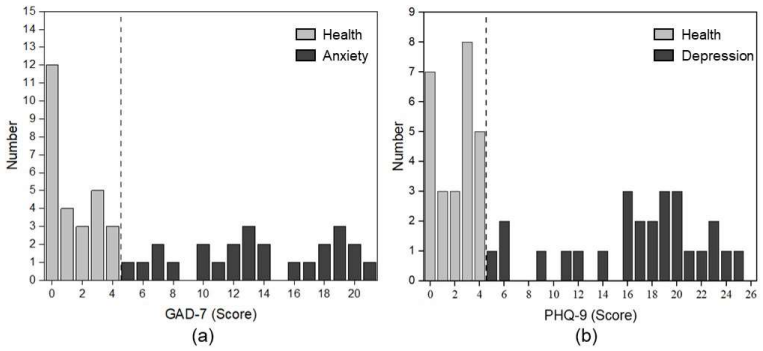

Figure 1: The scores distribution of the two typical scales with (a) for the anxiety scale GAD-7 and (b) for the depression scale PHQ-9. The vertical dashed lines indicate the threshold.

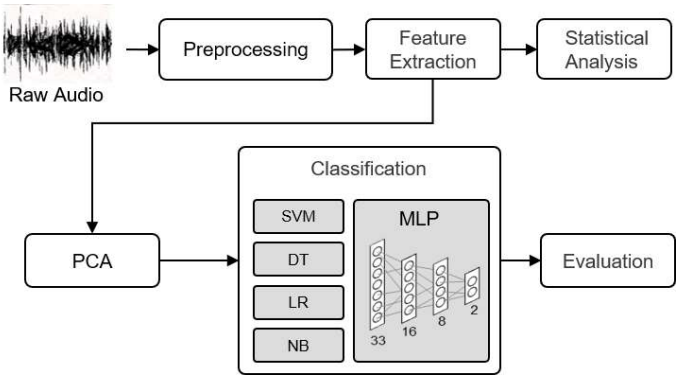

Figure 2: Visualization of the overall process of the proposed depression analysis and recognition method.

3.2 Feature Extraction

In this study, we extracted 1) prosodic feature, including voiced duration, unvoiced duration, fundamental frequency (F0), short-time energy (STE), zero crossing rate (ZCR) and sound pressure level (SPL); 2) voice-quality feature, including the center frequency and bandwidth of the first three formants (F1-F3), jitter and shimmer; and 3) spectral feature, including 13-dimensional Mel-frequency Cepstral Coefficients and its first and second derivatives (MFCC0-38), as described in Table 1 [17][18]. These features were first extracted frame-by-frame as LLD, which reflects local information. Then we calculated the mean and standard deviation of each prosodic and voice-quality feature and the mean of spectral feature as a 67-dimensional HSF to represent the global utterance-level features. Because the depression-related information hides in the state of the speaker, HSF retains overall characteristics and ignores the interference of transient variation. Each dimension feature was normalized to 0-1.

3.3 Statistical Analysis

Statistical analysis is performed on the acoustic features between groups to reveal the difference in depression-related pronunciation and further analyze the inducing effect. First, we used the Kolmogorov–Smirnov test to verify the normal distribution of samples. Then, Levene's test was used to test the homogeneity of variance. If satisfied, the Student's t-test was used; otherwise, Welch's t-test was used. All analyses were conducted with SPSS Statistics.

3.4 Dimension Reduction

PCA is used for dimension reduction in order to retain useful information and improve learning efficiency. The training set fits the covariance matrix and the test set transforms. To retain approximately 90% of cumulative contributions, the first 50% features with 33 dimensions were obtained for classification.

3.5 Classification

Four popular traditional classifies and one neural network model were built for depression recognition under the same conditions. The introduction and parameter settings of models are as follows:

- **Logistic Regression.** As one of the linear models, LR attempts to build a functional relationship between two or more predictor (independent) variables and one outcome (dependent) variable. This work used LIBLINEAR algorithm for optimization.
- **Support Vector Machine.** It can construct a hyperplane to discriminate two classes with the max distance. And it

Table 1: The description of acoustic features

Feature		Description
Prosodic Feature	Voiced Duration	The sound of the vocal cord vibration when the pronunciation
	Unvoiced Duration	The sound of the vocal cord does not vibrate when the pronunciation
	Fundamental Frequency	The frequency of the pitch in a polyphony, reflecting the pitch of the tone.
	Short-Time Energy	The sum of squares assigned to all speech signals in each frame.
	Zero Crossing Rate	The number of times a signal crosses the horizontal axis per frame
	Sound Pressure Level	The intensity of audio in air per frame, in dB.
Voice-quality Feature	First Three Formant Frequency and Bandwidth	The region of the spectrum in which energy is concentrated, reflecting the sound quality and physical properties of the vocal tract.
	Jitter	The variation of acoustic basic frequency between adjacent periods mainly, reflecting he degree of rough sound.
	Shimmer	The variation of acoustic amplitude between adjacent periods mainly, reflecting the degree of hoarseness.
Spectral feature	13-dimensional Mel-frequency Cepstral Coefficients + delta + delta delta	The short-term power spectrum of audio, reflecting the nonlinear characteristics of human ear frequency.

can manage linearly inseparable problems. We selected RBF kernel as the optimization function and used grid search to calculate the optimal parameters.

- **Decision Tree.** It can split data into binary categories using progressive iterations and establish the tree structure diagram. Each node represents a point at which the data are split, and the leaves at the end of the tree are the output variables.
- **Naive Bayes.** NB classifier is a kind of probabilistic classifiers based on applying Bayes' theorem. It assumes that all features are strongly independent. This model is easy to calculate and can be applied to reasonably large datasets.
- **Multilayer Perceptron.** With its feedback structure, MLP can adjust the weight of neurons and has a strong fitting ability [19]. In this work, the neuron number of input layer was 33, which was consistent with the output of feature dimension reduction. And the neuron number of hidden and output layers was set to 16, 8 and 2. Sigmoid activation function was used. Cross-entropy was selected as the loss function.

3.6 Evaluation

We used accuracy and F1 score as indexes to evaluate recognition performance. F1 score represents the harmonic mean of precision and recall, as given by:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1\ score = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

where TP/FP indicates true/false positives samples and TN/FN indicates true/false negatives samples.

4 RESULTS AND DISCUSSION

Here, we explored the induction effect of different speech tasks by statistical analysis and recognition performance, then compared the performance of MLP with traditional classifies.

4.1 Acoustic feature differences between groups

Figure 3 shows the distribution of significantly discriminative acoustic features between groups. As described in Table 1, these differences showed that the depressed group had a deep and slow voice, such as lower STE and SPL, which was consistent with clinical characteristics [6][7]. We further counted the number of features with significant differences in each speech task as the subgraph above. These significant differences are more prominent in the interview and picture description than in story reading and word reading tasks. We classified the interview and picture description as spontaneous pronunciation because the answers include the experience and understanding of the subjects. Spontaneous pronunciation is highly involved in cognitive and memory functions, which are impaired in depression as reported in [4][5]. We also classified the story reading and word reading tasks as mechanical pronunciation. We believe that mechanical pronunciation is difficult to arouse the cognitive and memory function due to the unfamiliarity with materials and the less interaction. From statistical analysis, spontaneous pronunciation induced more significantly discriminative acoustic features than mechanical pronunciation. However, due to the complexity of pronunciation and the correlation of acoustic features

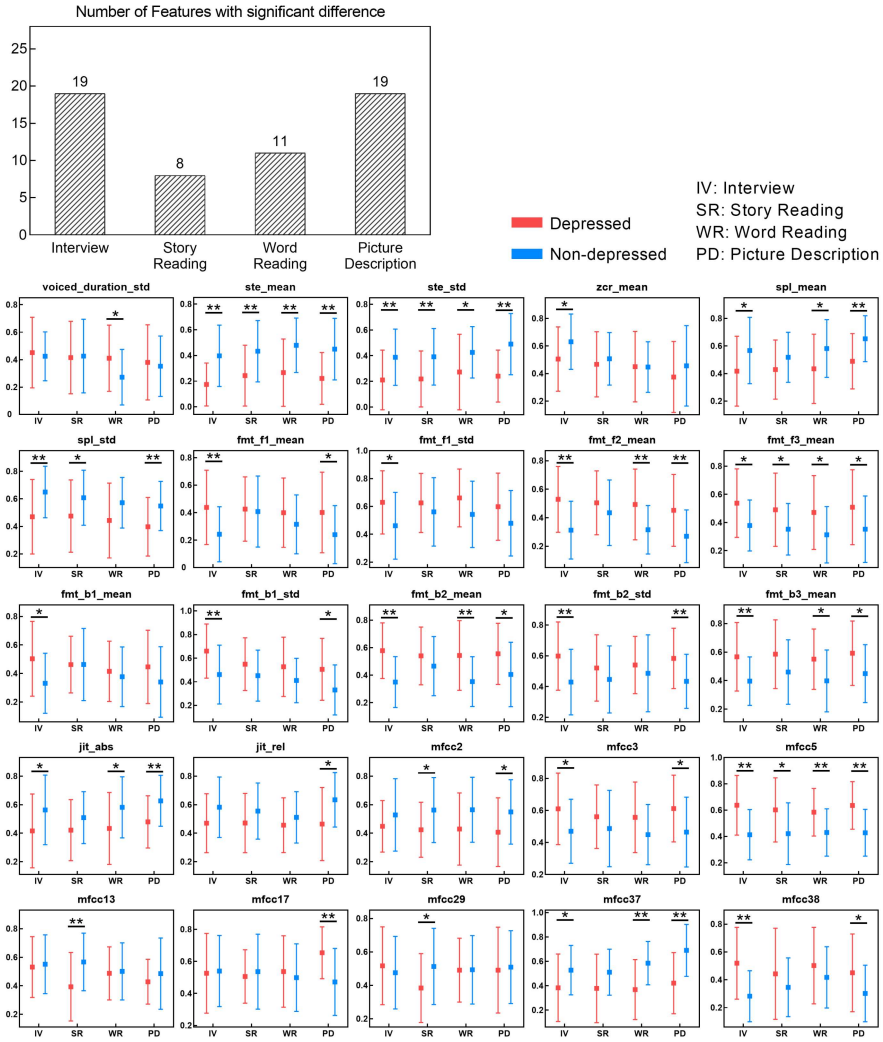

Figure 3: Results of statistical analysis for acoustic features between groups. * for p<0.05, ** for p<0.01.

[20], the recognition performance of different tasks needs to be further verified.

4.2 Recognition performance

We calculated the accuracy and F1 score of five classifiers under four speech tasks. As shown in Figure 4, MLP outperformed the recognition performance of LR, SVM, DT and NB under four tasks, indicating that MLP has a better fitting ability for acoustic features than traditional classifiers. Table 2 shows detailed results of MLP under four tasks. It can be seen that the recognition performance of interview and picture description under MLP also exceeded story reading and word reading, indicating that spontaneous pronunciation is more suitable for depression-related features than mechanical pronunciation, which is consistent with the statistical analysis results. In particular, MLP achieved the best recognition performance with accuracy 0.875 and average F1 score 0.855 under the picture description task. The result further improved the recognition performance by selecting speech tasks and improving the classifier and encouraged spontaneous pronunciation for future task design.

5 CONCLUSION

This work compared the induction effects of four speech tasks, then proposed an automatic depression recognition method from spontaneous pronunciation using machine learning. The results indicated that spontaneous pronunciation tasks (interview and picture description) induced more significantly discriminative acoustic

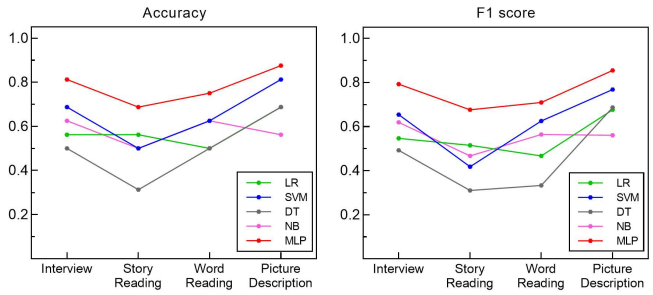

Figure 4: Accuracy and F1 score comparison of five classifiers under four speech tasks.

Table 2: Performance comparison of different speech tasks under MLP. ND for non-depression and D for depression

Speech Task	Accuracy	Precision	Recall	F1 score (D/ND)	F1 score (AVG)
Interview	0.813	0.667	0.800	0.727/0.857	0.792
Story Reading	0.688	0.667	0.571	0.615/0.737	0.676
Words Reading	0.750	0.500	0.750	0.600/0.818	0.709
Picture Description	**0.875**	**0.667**	**1.000**	**0.800/0.909**	**0.855**

features and achieved better recognition performance than mechanical pronunciation tasks (story reading and word reading). And MLP achieved the best recognition performance with accuracy 0.875 and average F1 score 0.855 under the picture description task than traditional classifiers, which has the potential as a complement to assist in the mass screening for depression. Meanwhile, this work provides a direction for follow-up research on task design and classifier building for automatic depression recognition.

ACKNOWLEDGMENTS

Research supported by the National Natural Science Foundation of China under Grant 81925020 and 81801786, and the General Program of Tianjin, China under Grant 19JCYBJC29200. The authors sincerely thank to the collectors and participants for providing the audio data for this study.

REFERENCES

[1] Maria Semkovska. 2021. Cognitive function and neurocognitive deficits in depression. The Neuroscience of Depression. (March 2021). 361-371. https://doi.org/10.1016/B978-0-12-817935-2.00021-0

[2] Global Health Data Exchange (GHDx). Institute of Health Metrics and Evaluation. Retrieved May 1, 2021 from http://ghdx.healthdata.org/gbd-results-tool?params=gbd-api-2019-permalink/d780dffbe8a381b25e1416884959e88b

[3] Ishara Madhavi, Sadil Chamishka, Rashmika Nawaratne, Vishaka Nanayakkara, Damminda Alahakoon, and Daswin De Silva. 2020. A Deep Learning Approach for Work Related Stress Detection from Audio Streams in Cyber Physical Environments. In 2020 25th IEEE International Conference on Emerging Technologies and Factory Automation (ETFA). IEEE, Vienna, Austria, 929–936. https://doi.org/10.1109/ETFA46521.2020.9212098

[4] T. Jannini, L. Longo, F. Marasco, M. Di Civita, C. Niolu, A. Siracusano, and G. Di Lorenzo. 2021. Cognitive function and metabolic syndrome in unipolar and bipolar depression: A pilot study. European Psychiatry. 64. S1(April 2021). S82–S82. https://doi.org/10.1192/j.eurpsy.2021.246

[5] Nicholas Cummins, Stefan Scherer, Jarek Krajewski, Sebastian Schnieder, Julien Epps, and Thomas F. Quatieri. 2015. A review of depression and suicide risk assessment using speech analysis. Speech communication. 71, C(July 2015). 10–49. https://doi.org/10.1016/j.specom.2015.03.004

[6] Bethany Little, Ossama Alshabrawy, Daniel Stow, I. Nicol Ferrier, Roisin McNaney, Daniel G. Jackson, Karim Ladha, Cassim Ladha, Thomas Ploetz, and Jaume Bacardit. 2021. Deep learning-based automated speech detection as a marker of social functioning in late-life depression. Psychological medicine. 51, 9(July 2021): 1441–1450. https://doi.org/10.1017/S0033291719003994

[7] Brian Stasak, Julien Epps, and Roland Goecke. 2019. Automatic depression classification based on affective read sentences: Opportunities for text-dependent analysis. Speech Communication. 115, (December 2019). 1–14. https://doi.org/10.1016/j.specom.2019.10.003

[8] Jonathan Gratch, Ron Artstein, Gale Lucas, Giota Stratou, Stefan Scherer, Angela Nazarian, Rachel Wood, Jill Boberg, David DeVault, and Stacy Marsella. 2014. The distress analysis interview corpus of human and computer interviews. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14). Reykjavik, Iceland, 3123–3128.

[9] Michel Valstar, Björn Schuller, Kirsty Smith, Florian Eyben, Bihan Jiang, Sanjay Bilakhia, Sebastian Schnieder, Roddy Cowie, and Maja Pantic. 2013. Avec 2013: the continuous audio/visual emotion and depression recognition challenge. In Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge. Association for Computing Machinery, New York, USA, 3–10. https://doi.org/10.1145/2512530.2512533

[10] Takaya Taguchi, Hirokazu Tachikawa, Kiyotaka Nemoto, Masayuki Suzuki, Toru Nagano, Ryuki Tachibana, Masafumi Nishimura, and Tetsuaki Arai. 2018. Major depressive disorder discrimination using vocal acoustic features. Journal of Affective Disorders. 225. (January 2018). 214–220. https://doi.org/10.1016/j.jad.2017.08.038

[11] Michel Valstar, Jonathan Gratch, Björn Schuller, Fabien Ringeval, Denis Lalanne, Mercedes Torres Torres, Stefan Scherer, Giota Stratou, Roddy Cowie, and Maja Pantic. 2016. Avec 2016: Depression, mood, and emotion recognition workshop and challenge. In Proceedings of the 6th international workshop on audio/visual emotion challenge. Association for Computing Machinery, New York, USA, 3–10. https://doi.org/10.1145/2988257.2988258

[12] Gauri Deshpande, Venkata Subramanian Viraraghavan, Mayuri Duggirala, and Sachin Patel. 2019. Detecting emotional valence using time-domain analysis of speech signals. In 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), IEEE, Berlin, Germany, 3605–3608. https://doi.org/10.1109/EMBC.2019.8857691

[13] Pan Wei, Wang Jingying, Liu Tianli, Liu Xiaoqian, Liu Mingming, Hu Bin, and Zhu Tingshao. 2018. Depression recognition based on speech analysis. Chinese Science Bulletin. 63. 20(July 2018). 2081–2092. https://doi.org/10.1360/N972017-01250

[14] Xingchen Ma, Hongyu Yang, Qiang Chen, Di Huang, and Yunhong Wang. 2016. DepAudioNet: An Efficient Deep Model for Audio based Depression Classification. In Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge. Association for Computing Machinery, New York, USA, 35–42. https://doi.org/10.1145/2988257.2988267

[15] Hanshu Cai, Yiwen Gao, Shuting Sun, Na Li, Fuze Tian, Han Xiao, Jianxiu Li, Zhengwu Yang, Xiaowei Li, and Qinglin Zhao. 2022. MODMA dataset: a Multi-modal Open Dataset for Mental-disorder Analysis. Scientific Data. 9. 1(April 2022). 1-10. https://doi.org/10.1038/s41597-022-01211-x

[16] Bo Liu, Hongli Chang, Kang Peng, and Xuenan Wang. 2022. An End-to-End Depression Recognition Method Based on EEGNet. Front Psychiatry. 13. 864393 (March 2022). https://doi.org/10.3389/fpsyt.2022.864393

[17] Sanne Koops, Sanne G. Brederoo, Janna N. de Boer, Femke G. Nadema, Alban E. Voppel, and Iris E. Sommer. 2022. Speech as a Biomarker for Depression. CNS & Neurological Disorders Drug Targets. (February 2022). https://doi.org/10.2174/1871527320666211213125847

[18] Sara C. Keen, Karan J. Odom, Michael S. Webster, Gregory M. Kohn, Timothy F. Wright, and Marcelo Araya-Salas. 2021. A machine learning approach for classifying and quantifying acoustic diversity. Methods in Ecology and Evolution. 12. 7 (July 2021). 1213–1225. https://doi.org/10.1111/2041-210X.13599

[19] Abdullah Al Mamun Sardar, Sanzidul Islam, and Touhid Bhuiyan. 2021. A Review on Automatic Speech Emotion Recognition with an Experiment Using Multi-layer Perceptron Classifier. Soft Computing Techniques and Applications. 1248. (November 2020). 381–388. https://doi.org/10.1007/978-981-15-7394-1_36

[20] Yikang Wang and Hiromitsu Nishizaki. 2022. Combination of Time-domain, Frequency-domain, and Cepstral-domain Acoustic Features for Speech Commands Classification. (June 2022). https://doi.org/10.48550/arXiv.2203.16085