Investigation of the Accuracy of Depression Prediction Based on Speech Processing

Gábor Kiss, Attila Zoltán Jenei
Department of Telecommunications and Media Informatics
Budapest University of Technology and Economics
Budapest, Hungary
Email: kiss.gabor@tmit.bme.hu, ja1504@hszk.bme.hu

***Abstract*—The present study investigates the accuracy of prediction of depression based on speech processing. Depression is one of the most widespread psychiatric disorders, but early detection of depression is difficult. The Beck Depression Inventory II (BDI) is a self-assessment questionnaire and can accurately predict the severity of depression. BDI is most often used for early detection. There is no known objective biomarker for depression, but the state alters the speech of the individual suffering from it, providing an opportunity for speech-based detection. In the current study, we investigated the accuracy of prediction of depression severity based on speech signal processing, and how accurately it can predict the severity of depression compared to the Beck Depression Inventory.**

***Keywords*—BDI; depression prediction; HAM-D; speech processing; SVR**

I. Introduction

Many diseases affect the speech production system as well as the acoustic characteristics of speech. For this reason, it is possible to view speech as an objective biomarker, which can warn us in the early appearing of several disorders such as depression [1], [2], Parkinson's disease [3], [4], laryngeal disorders [5], [6].

Depression is a psychiatric disorder. The World Health Organization (WHO) estimated the number of depressed patients at 350 million in 2012 [7]. According to WHO forecasts [8], by 2030, depression will be one of the three most serious diseases worldwide with HIV / AIDS and heart disease.

Patients with depression suffer a significant deterioration in their quality of life and may even be unable to work, depending on the severity of the symptoms. This is a major economic problem for society [9]. In addition, it also increases the risk of suicide [10].

Despite the fact that the number of people suffering from depression is very high, it is the responsibility of a small number of qualified specialists (psychiatrists) to make the diagnosis. For this reason, any objective system for the support of diagnosis is very useful. The fact that depression influences speech allows us to view human speech as objective biomarker to support the diagnosis.

TABLE I. Depression Severity categories of HAM-D and BDI

Category	HAM-D score	BDI score
not depressed	0-7	0-13
mild depression	8-13	14-19
moderate depression	14-18	20-28
severe depression	19-22	29-63
very severe depression	23-48	

Numerous studies have shown that speech processing can also be used to successfully identify depression and predict the severity of depression [1], [2].

In this study, we investigate the accuracy of prediction of the severity of depression based on speech processing and we compare the achieved results with the accuracy of the Beck Depression Inventory II (BDI) questionnaire [11], which is currently one of the most common method for early warning of depression in clinical practice.

II. Methods

A. Depression Severity Scales

Several scales are used to describe the severity of depression. In this research, we used one of the two most commonly used scales, the BDI [11] and the Hamilton Depression Rating Scale (HAM-D) [12]. Both scales are commonly used in clinical practice, but there are significant differences between them. While BDI is a self-assessment questionnaire that focuses mainly on negative self-esteem symptoms, the HAM-D score can only be determined by a psychiatrist and prefers neuro-vegetative symptoms. Therefore, the HAM-D score is considered to be the more objective of the two. However, it is more difficult to determine the HAM-D score, not only because only a psychiatrist can determine the score, but because the BDI questionnaire can be completed within 5-10 minutes, whereas the HAM-D score can be determined within 20-30 minutes. It is important to note that the BDI and HAM-D scales measure the severity of depression in different ranges, as shown in Table I.

B. Database

The Hungarian Depressed Speech Database was used in the research [2].

Project no. K128568 has been implemented with the support provided from the National Research, Development and Innovation Fund of Hungary, financed under the K_18 funding scheme.

TABLE II. DESCRIPTION OF THE USED DATABASE

	Dataset A	Dataset B	Overall
Speech Recordings	158 103 female 55 male	22 16 female 6 male	180 119 female 61 male
Mean of BDI	15.7	28.2	17.2
Variance of BDI	13.2	9.1	13.4

The speech recordings were collected by the Psychiatric and Psychotherapeutic Clinic of Semmelweis University, Budapest. The recordings were recorded at 44,1 kHz with 16-bit sample rate. During the collection of speech recordings, speakers covered different degrees of severity of depression, from healthy to severely depressed. Spontaneous and read speech were recorded, from which we used only read samples. Speakers had to read a short story in Hungarian ("North Wind and the Sun"). The text was about 10 sentences long. Speech recordings were collected only from subjects who did not suffer from any other disorder, except depression. The database contains exactly one recording from each speaker. The BDI score was recorded in all cases and for 22 subjects of them additionally the HAM-D score was noted. Thus, the database was split into two sets: "Dataset A", which contains speech recordings where only the BDI score is given and used for training purposes. “Dataset B”, which contains speech recordings where both BDI and HAM-D scores are given and used for testing purposes. The split of the database and the BDI distribution of each dataset is shown in the Table II.

C. *Preprocessing and Segmentation*

Amplitude was normalized to peak in each speech recording to eliminate possible differences in recording settings (different gain or microphone distance). Each speech recording was segmented and labelled on phoneme level using an automated forced alignment segmentation method [13].

D. *Acoustic Features*

A total of 204 features were extracted from each speech recording. The following low-level descriptors (LLDs) were extracted: jitter, shimmer, fundamental frequency (f0), intensity, 21 mel- frequency band energy values, first and second formant frequency values (F1 and F2). Each LLD was computed at 10-ms data frames using Praat software [14].

The mean, standard deviation and percentile ranges of 1%, 2.5% and 5% of the above mentioned LLDs were calculated together for all sounds, for all vowels and for vowels /e/ separately (vowel /e/ was the most common in the tale). The final derived features were normalized between -1 and 1 for the feature vector before passing them to the machine learning algorithm.

E. *Model training*

Support Vector Regression [15] method with rbf kernel function was used for regression purposes. BDI was used as the target variable. Due to the limited amount of data it is necessary to select the appropriate model. The feature vector was selected by fast forward selection method, and kernel hyper parameters (cost and gamma) were selected by grid search in the range between 2^{-5} and 2^{5}. Selection was performed using leave-one-out cross validation (LOOCV) method on “Dataset A”. Separate model was selected and trained for females and males.

The grid search selected cost = 2^4 and gamma = 2^{-4} as optimal for both sexes. However, minor differences can be observed in the selection of gender-specific feature vectors. In case of female model, the selected feature vector contained 13 feature, in which the derived characteristics of mel- frequency band energy values, intensity, and shimmer were found. In case of male model, the feature vector contained only 11 features although it contains derived features of the same LLDs.

F. *Testing*

“Dataset B” was used for testing purposes. In case of testing we compared the predicted BDI scores with the original BDI scores and if it was possible with the HAM-D scores. However, as mentioned earlier, the two scales work with different ranges, which necessitates the transformation of the HAM-D scale. We solved this by fitting the endpoints of the categories, and using linear interpolation within each category. Thus, we could compare the predicted BDI scores and the transformed HAM-D scores. In addition, the original BDI scores and the transformed HAM-D scores can be compared. For the evaluation, we examined the root mean square error (RMSE) value, the mean absolute error (MAE) value, the Pearson correlation coefficients of the predicted scores and the reference scores.

III. RESULTS

A. *Accuracy of the Acoustic Model based on the Original BDI score*

Using “Dataset A”, the best achieved result was 7.2 RMSE using LOOCV method. The same model was applied on “Dataset B”, achieving 9.4 RMSE using "normal" testing. The relationship between the original and predicted BDI scores is shown in Figs. 1 and 2 for these two datasets.

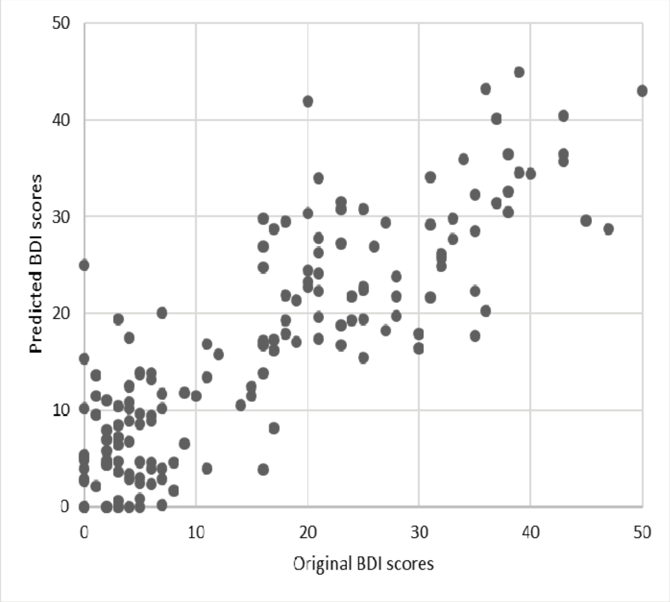

Fig. 1. The relationship between the original and predicted BDI scores in case of “Dataset A” using LOOCV

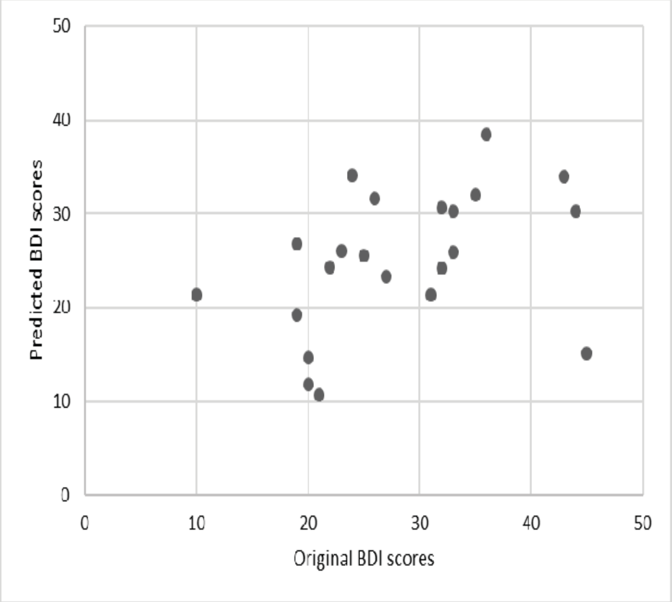

Fig. 2. The relationship between the original and predicted BDI scores in case of "Dataset B" using "normal" testing

Further descriptive characteristics of the two tests are shown in the Table III. It can be stated that the results on "Dataset A" outperformed the results on "Dataset B". There can be many reasons explaining the results, but the most obvious factor is that the two datasets differ in the distribution of speakers according to their severity of depression.

B. *Comparison of the Accuracy of the Acoustic Model and BDI Questionnaire*

HAM-D scores were available for "Dataset B", which can be considered a more accurate and objective measure of the severity of depression. For this reason, we examined the relationship between the predicted BDI scores and the transformed HAM-D scores in this case (this was not possible for "Dataset A" because HAM-D scores were not available for those recordings). Furthermore, the relationship between the original BDI scores and the transformed HAM-D scores was investigated. The results of these two examinations are shown in Figs. 3 and 4.

TABLE III. ACCURACY OF THE ACOUSTIC MODEL ON "DATASET A" AND ON "DATASET B"

Dataset	RMSE	MAE	Pearson cor.
Dataset A	7.2	5.7	0.84
Dataset B	9.4	8.0	0.41

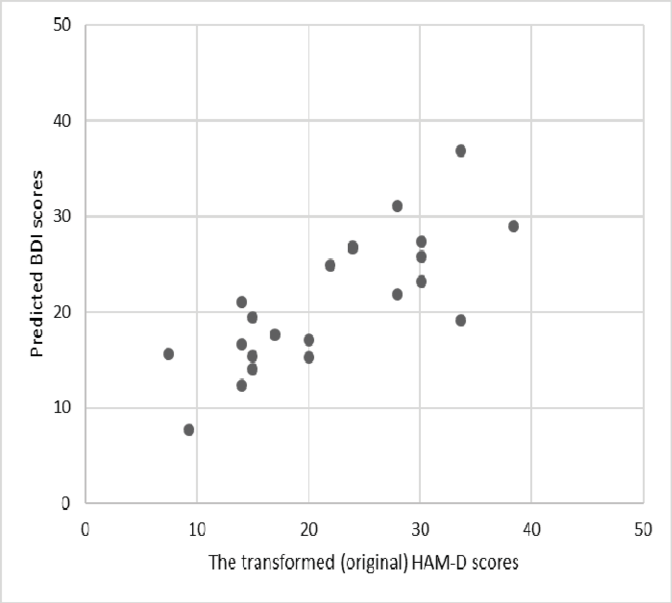

Fig. 3. The relationship between the transformed HAM-D scores and the predicted BDI scores based on speech processing in case of "Dataset B" using normal testing

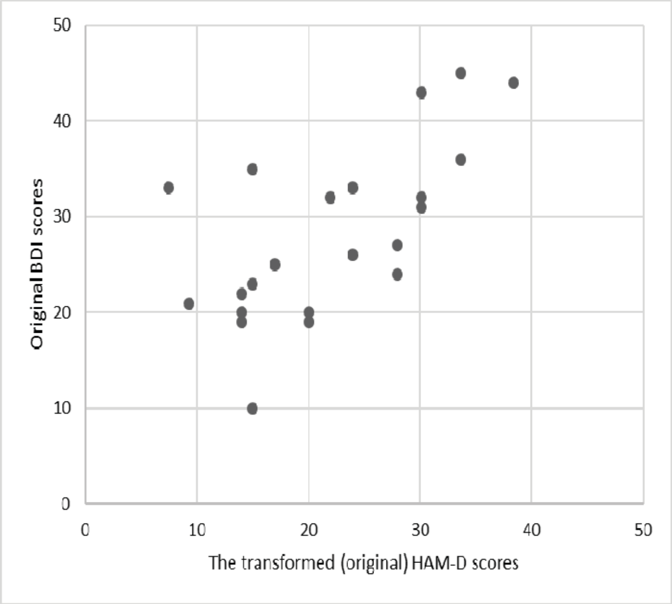

Fig. 4. The relationship between the transformed HAM-D scores and the original BDI scores in case of "Dataset B"

Further descriptive characteristics of these two examinations are shown in the Table IV.

By examining the results (Table III. row 2 and Table IV. row 1), it can be stated that the predicated scores of the acoustic model are closer to the transformed HAM-D scores (RMSE: 5.4, MAE: 4.3) than to the original BDI scores (RMSE: 9.4, MAE: 8.0), that means, the error rate was lower when the predicted scores were compared with the transformed HAM-D scores.

TABLE IV. Comparison of the Accuracy of the Acoustic Model and BDI Questionnaire based on the transformed HAM-D scores

	RMSE	MAE	Pearson cor.
Acoustic model	5.4	4.3	0.78
Original BDI scores	9.5	7.2	0.65

This is significant because the model was trained based on BDI scores, meaning that the acoustic model was probably able to learn the characteristics of the depressed state itself, despite the possible bias of the BDI questionnaire.

Furthermore, it can be stated that the acoustic model was able to predict the severity of depression with a lower error rate (RMSE: 5.4, MAE: 4.3) than the BDI questionnaire (RMSE: 9.5, MAE: 7.2) compared to the transformed HAM-D score.

IV. Conclusion

This study investigated the accuracy of predicting the severity of depression based on speech processing. For model training, BDI scores were used as target to describe the severity of depression. Both BDI and HAM-D scores were used for testing, but HAM-D scores can be considered more accurate and objective measurement of the severity of depression.

Our examinations have shown that it is possible to accurately predict the severity of depression based on speech processing and prediction scores are closer to HAM-D scores than BDI scores. This suggests that the model was actually able to learn the speech characteristics of depression and its effect on speech depending on its severity. This is particularly significant given the fact that only BDI scores were available for model training. Furthermore, it was shown that the speech-based model predicted the severity of depression more accurately than the BDI questionnaire at least on the given dataset.

Of course, we are aware that the size of the test set was limited, so we would definitely like to continue our investigation by examining more speech recordings where both BDI and HAM-D scores are available. However, the results achieved so far give reason to believe that speech processing systems can be more accurate than the BDI questionnaire, which is currently one of the most common methods for early recognition of depression in clinical practice.

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 Communication 71, pp. 10–49, 2015.

[2] G. Kiss and K. Vicsi, “Mono-and multi-lingual depression prediction based on speech processing,” International Journal of Speech Technology, 4, pp. 919-935, 2017.

[3] J. R. Orozco-Arroyave, F. Hönig, J. D. Arias-Londoño, J. Vargas-Bonilla, S. Skodda, J. Rusz and E. Nöth, "Voiced/unvoiced transitions in speech as a potential bio-marker to detect Parkinson's disease," in INTERSPEECH 2015 pp. 95-99, Dresden, Germany, 2015.

[4] J. Vásquez-Correa, J. R. Orozco-Arroyave, and E. Nöth, “Convolutional neural network to model articulation impairments in patients with Parkinson’s disease,” in INTERSPEECH 2017 pp. 314–318, Stockholm, Sweden, 2017.

[5] Y. Liu, T. Lee, P. C. Ching T. K. T. Law and K. Y. S. Lee, “Acoustic assessment of disordered voice with continuous speech based on utterance-level ASR posterior features,” in INTERSPEECH 2017 pp. 2680–2684, Stockholm, 2017.

[6] M. G. Tulics and K. Vicsi, “Phonetic-class based correlation analysis for severity of dysphonia,” in 8th IEEE Conference on Cognitive Infocommunications (CogInfoCom2017) pp. 21–26. 2017.

[7] M. Marcus, M. T. Yasamy, M. V. van Ommeren, D. Chisholm and S. Saxena. ”Depression: A global public health concern,” WHO Department of Mental Health and Substance Abuse, 1, 6-8, 2012.

[8] C. D. Mathers and D. Loncar, ”Projections of global mortality and burden of disease from 2002 to 2030,” Plos med, 3(11), e442, 2006.

[9] J. Olesen, A. Gustavsson, M. Svensson, H. U. Wittchen, B. Jönsson, “The economic cost of brain disorders in Europe,” European Journal of Neurology, 19(1), 155-162, 2012.

[10] K. Hawton, C. C. i Comabella, C. Haw, K. Saunders, “Risk factors for suicide in individuals with depression: a systematic review,” Journal of Affective Disorders, 147(1-3), 17-28, 2013.

[11] A. T. Beck, R. A. Steer, R. Ball and W. F. Ranieri, “Comparison of beck depression inventories -IA and -II in psychiatric outpatients,” Journal of Personality Assessment 67, pp. 588–597, 1996.

[12] H. Hamilton, “HAMD: a rating scale for depression,” Neurosurg. Psych. 23, 56–62, 1960.

[13] D. Sztahó, G. Kiss, L. Czap, and K. Vicsi, ”A computer-assisted prosody pronunciation teaching system.,” in WOCCI, pp. 45-49, 2014.

[14] P. Boersma, ”Praat, a system for doing phonetics by computer.” Glot international, 5(9/10), pp.341-345, 2002.

[15] H. Drucker, C. J. Burges, L. Kaufman, A. J. Smola, V. Vapnik, “Support vector regression machines,” In Advances in Neural Information Processing Systems (pp. 155-161), 1997.