https://doi.org/10.3311/PPee.15958

Periodica Polytechnica Electrical Engineering and Computer Science, 65(3), pp. 227–234, 2021

Severity Estimation of Depression Using Convolutional Neural Network

Attila Zoltán Jenei[1*], Gábor Kiss[1]

[1] Department of Telecommunications and Media Informatics, Faculty of Electrical Engineering and Informatics, Budapest University of Technology and Economics, H-1117 Budapest, Magyar tudósok körútja 2, Hungary
[*] Corresponding author, e-mail: jenei@tmit.bme.hu

Received: 16 March 2020, Accepted: 25 March 2021, Published online: 29 June 2021

Abstract

In the present study, we attempt to estimate the severity of depression using a Convolutional Neural Network (CNN). The method is special because an auto- and cross-correlation structure has been crafted rather than using an actual image for the input of the network. The importance to investigate the possibility of this research is that depression has become one of the leading mental disorders in the world. With its appearance, it can significantly reduce an individual's quality of life even at an early stage, and in severe cases, it may threaten with suicide. It is therefore important that the disorder be recognized as early as possible. Furthermore, it is also important to determine the disorder severity of the individual, so that a treatment order can be established. During the examination, speech acoustic features were obtained from recordings. Among the features, MFCC coefficients and formant frequencies were used based on preliminary studies. From its subsets, correlation structure was created. We applied this quadratic structure to the input of a convolutional network. Two models were crafted: single and double input versions. Altogether, the lowest RMSE value (10.797) was achieved using the two features, which has a moderate strength correlation of 0.61 (between estimated and original).

Keywords

BDI estimation, convolutional neural network, depression, MFCC

1 Introduction

Depression is one of the most common psychiatric disorder affecting more than 264 million people worldwide [1, 2]. Its symptoms are wide, which makes it difficult to make an accurate diagnosis. The detection rates are low and patients are therefore not receiving appropriate treatment.

Exact causes of depression are not yet known however; the physiological symptom of depression is most often a form of dysfunction of the cortical limbic system [3].

Depending on the severity of the depression, the individual may find it difficult to cope with personal and/or social activities [4]. In addition, worsening of depression may increase the risk of suicide [5]. Finally, diagnosing the disease requires specialist knowledge, which falls on a small group of doctors. Therefore, developing a high-accuracy diagnostic tool to help doctors work has a primary focus. This will help the patient to be diagnosed earlier and receive adequate treatment.

There is large amount of research works that explore the applicability of several biomarkers in the recognition of depression. Speech is such a biomarker. It can indicate not only depression but also many other illnesses, such as Parkinson's disease [6] or dysphonia [7]. Speech provides an opportunity to develop effective, non-invasive diagnostic tools that can assist professionals in their work [8, 9].

Automatic detection of depression is a recent area of research based on speech production, primarily due to the increasing depressed speech databases and the advancement of information technology [10].

A wide range of speech acoustic features is available for an objective description of speech production. A subset of these are widely used to recognize and estimate the severity of depression [11–13]. Descriptive features are generally divided into prosodic and spectral sets [14]. The former includes pitch, speech rate, jitter and shimmer while the latter contains mel-band energy values, MFCCs (mel-frequency cepstral coefficients), formants and their bandwidths.

Numerous studies have shown that acoustic features calculated from speech correlate with the severity of depression. However, it is still an open research question to demonstrate the effectiveness of acoustic features in separating different levels of depression severity as well [15–17].

Cite this article as: Jenei, A. Z., Kiss, G. "Severity Estimation of Depression Using Convolutional Neural Network", Periodica Polytechnica Electrical Engineering and Computer Science, 65(3), pp. 227–234, 2021. https://doi.org/10.3311/PPee.15958

One of the challenges of AVEC in 2013 was to assess depression severity [12] measured by Beck Depression Inventory (BDI-II) scale questionnaire. Training and validation sets were provided with 50–50 samples and their associated BDI values. Challenge participants had to estimate these BDIs with as little error as possible.

As baseline, support vector machine regression results with linear kernel were provided for the test set: 14.12 RMSE (Root mean square error) and 10.35 MAE (Mean absolute error). For this, the features of the recordings were extracted using basic segmentation. The best results were achieved by using 20 seconds long non-overlapping segments, which were averaged on the full recording.

The winner of the challenge was Williamson et al. [18], who achieved 7.42 RMSE and 5.75 MAE as the best result on development set (and 0.80 of Pearson correlation). A Gaussian Mixture Model was used on a correlation matrix containing formants and delta-mel-cepstral coefficients.

In study [19], Lang He and his colleague conducted an investigation on AVEC2013 audio recordings. They have created their own models and deep convolutional neural networks (DCNN) to estimate severity. The best result on the development set was obtained with DCNN spectrogram model (RMSE = 9.122, MAE = 7.537). Similarly, using the test set, DCNN spectral-based model performed best (RMSE = 10.456, MAE = 8.483).

Similar correlation structure was used in our research as in Williamsons' work. However, while they used the eigenvalues of the structure for regression, the structure itself was applied here as an input image to a convolutional neural network to predict the severity of depression according to BDI-II. The advantage of this method is that the machine learning process also has the task of properly processing the correlation matrix. As a novelty, we were also able to test the procedure on speech samples from 91 patients, which is considered large among datasets internationally.

In the second Section, the speech database is presented. In the third Section, we present the BDI scale, the applied low-level features, the method of calculating the correlation matrix and the convolutional neural networks itself, as well as the evaluation methods. In the fourth Section, the results are detailed. Section 5 summarizes the main findings of the research and further plans briefly.

2 Hungarian depression speech database

The speech dataset includes 91 recordings from depressed people (DE) and 91 from healthy ones (HC). This is a constantly expanding database. Whose recordings were collected by the members of the Laboratory of Speech Acoustics. Recordings were collected from healthy subjects without any known disease that would affect speech production. Depressed patients included in the study had not been diagnosed with any other speech-related disorder.

Recordings from depressed patients were collected at Psychiatry and Psychotherapy Clinic of Semmelweis University. 21 mild, 32 moderate and 38 severe, (overall 91 depression) cases were included in the severity categories defined by BDI-II (Beck Depression Inventory-II) [20] in Table 1. 88 % of the depressed subjects (80 cases) can be found between 20 and 40 BDI score (moderate or severe depression).

The individuals had to read a phonetically rich tale ("The North Wind and the Sun"), with approximately 1-minute duration. Recordings were made in a quiet room with a clip-on microphone at 44.1 kHz sampling rate.

3 Methods

Fig. 1 shows the process developed in the research. Acoustic features were obtained from speech samples, from which a 2D auto- and cross-correlation structure was constructed. This was given to the convolutional network after separated to training and testing sets.

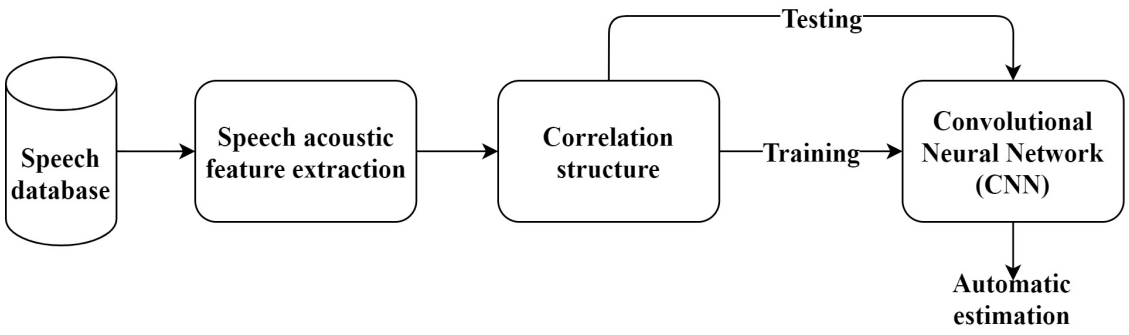

Fig. 1 Examination process elements: speech recordings, extracted acoustic features, correlation structure, CNN models with training and testing cycle.

3.1 Scales of depression severity

Two scoring systems are the most commonly used to estimate the severity of depression. These are the Beck Depression Inventory (BDI) and the Hamilton Depression Rating Scale (HAM-D).

The BDI is a 21-item self-assessment questionnaire, created in 1961 [21]. In 1978 and 1996, it underwent a review to form the BDI-II. The questionnaire was translated into several languages and it is commonly used in clinical practice. A shorter version consists of seven elements and is available as BDI-FS. In this study, BDI-II scale was used, because it was originally recorded by the doctor examining the patient. This questionnaire is also faster to complete and more common in practical application than HAM-D.

Each item has a score between 0 and 3, summing up to maximum 63 on the questionnaire. The severity categories for depression on BDI scale are shown in Table 1.

3.2 Extracted speech acoustic characteristics

In advance, sound recordings were normalized to peak amplitude. Then, with the help of the Praat [22], the following features were obtained.

MFCC: The MFCC values were calculated as the discrete cosine transform of the 27 mel-band energy values, from which the first 14 coefficients were used.

Formant frequency: A broad peak, or local maximum frequency in the spectrum. In the research, the first three-formant frequency values were calculated, hereinafter referred to as F1, F2, and F3.

The features were extracted in 10 millisecond steps with a 50 millisecond-sliding window. The MFCC coefficients were calculated throughout the whole speech sample and the formant frequencies in the voiced sections. Finally, 14 MFCC and 3 formant frequency time-like vectors were available for one subject. Later on, we refer this time-like vectors as feature signals.

3.3 Auto- and cross correlation structure

The illustration of the correlation structure is showed in Fig. 2 for feature signals 1 to n. An arbitrary cell holds the correlation of the two related feature signals. In this case, the matrix would be of size $n \times n$ (left side of Fig. 2). Instead of one correlation value, we created a submatrix that contains the correlation coefficient values along with the displacement in the feature signals (right side of Fig. 2).

The submatrices in the main diagonal are filled with autocorrelation, while the submatrices in the side diagonals are filled with cross-correlation coefficients.

The feature signals were shifted k–1 times (k–1 additional signal), so the size of the submatrices is $k \times k$ (by including the original signals). The size of the total correlation structure is ($n \times k$) x ($n \times k$).

In the submatrices, the first cell still has the correlation coefficient of the original signals (row 1, column 1). In a given row i and column j addressed cell, we put the correlation coefficient of two feature signals where the first feature signal was shifted by i–1 times and the second feature signal was shifted by j–1 times. For example, if $i = 5$, then the first feature signal was shifted by 4 times with the degree of displacement.

In this study, 1 and 8 were used for the degree of displacement and 10, 20 were applied for the k (later referred as offset). From the feature groups the first 14 MFCCs, the first 7 MFCCs and the first 3-formant frequencies were used. These features and values come from the [23] preliminary research where a larger number of features were examined for classification.

3.4 Convolutional Neural Network (CNN)

The CNN was created in Python using TensorFlow 1.8.

Two structures are presented in this study. The first one has one input layer (Fig. 3) while the second one has two (Fig. 4). In other words, the second CNN can handle two correlation structures from one person at the same time.

The first network contains two convolutional layers with 32 non-overlapping filters respectively with ReLU activation functions. The size of the first convolutional kernel

Table 1 Depression severity categories in BDI-II scale

	BDI-II	Number of patients
Normal	0–13	0
Mild depression	14–19	21
Moderate depression	20–28	32
Severe depression Very severe depression	29–63	38

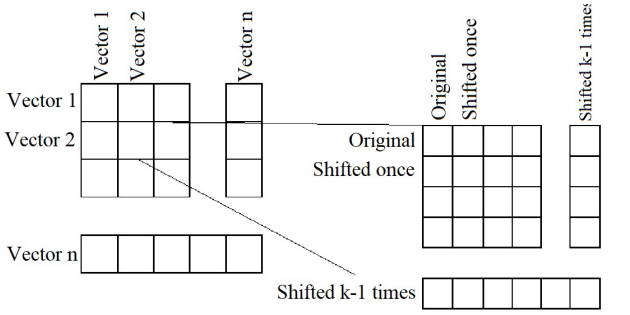

Fig. 2 The full correlation matrix from n features (left) and a submatrix of two features with k–1 displacement (right).

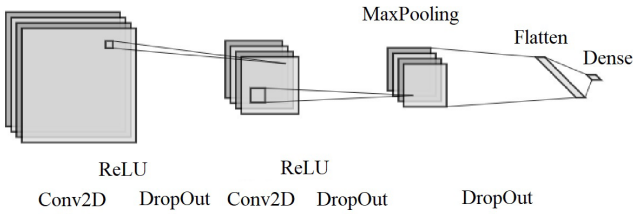

Fig. 3 Implemented convolutional network for one correlation structure (single channel model).

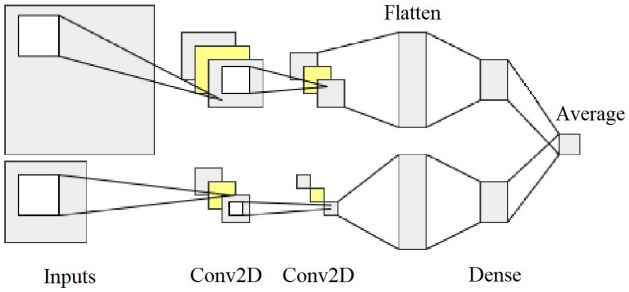

Fig. 4 Implemented convolutional network for two correlation structures (double channel model).

was adjusted to the size of the submatrix shown above. It was 10 × 10 for k = 10 and 20 × 20 for k = 20. The size of the second convolutional kernel was set to get the output size of 2 × 2. It was followed by a max-pooling layer with 2 × 2 pool size. Dropout regulation was applied after the first three layers, which ignored 25 % of the neurons during the training process. The flatten layer arranged the 32 filters value into a row vector. At the end, a dense (fully connected neural network) layer was implemented that served the estimation with a linear activation function. The input neuron number of the dense layer was 32 and the output was 1 (output 1 value).

The second network was designed to get two input correlation matrices from different features from one person. The first matrix was crafted from the MFCC values, and the second one from formants only. From here, both channels followed the construction of the first network to the dense layer, except that they are not contained max pooling layer. This is due to avoid excessive dimensional reduction. Furthermore, the second convolutional layer was applied with 64 filters instead of 32. The input size of the dense layer is 1 × 64 while the output is one number (normalized BDI value) in each of channels that was linked together by an average (merge) layer. To put it simply, the first model was applied twice with small modifications (one channel for MFCCs and one for formants) and then the outputs were averaged as the estimated (normalized) BDI value.

3.5 Implementation of evaluation

The values in the correlation structure were normalized between -1 and 1 while its associated BDI was normalized between 0 and 1.

To preprocess the correlation structure, all 182 samples were randomly shuffled to distribute the healthy and depressed sample as evenly as possible.

For training and testing the convolutional networks, full cross-validation was performed (leave one out cross-validation, LOOCV) that separated one element into a test set while the remainder served as training samples, iterating through all samples in the dataset. In this case, 182 training/testing process was performed (since 182 samples were used). The training and test sets were always disjoint and there was no overlap between them.

ADAM optimization algorithm was used with default parameters by Keras [23] to optimize learning rate. Mean squared error was used as cost function.

For training phase, the batch size was set to the number of training samples (batch size: 181). So in each epoch, the network could see the whole training set at once. Multiple epoch numbers were chosen: 25, 50, 75, 100, 150, 250, 500.

MFCC features were used at a displacement rate of 8 and formants at a displacement rate of 1 based on [24] preliminary study.

RMSE was used to determine the error between the original and estimated BDI values [25].

As a first approach, RMSE values were obtained by the first 7 and 14 MFCC feature signals separately in the network of Fig. 3 using two different offset (10 and 20).

The offset with lowest RMSE was selected and extended later on with a set of formant frequencies. This has been examined in the model of Fig. 4.

As additional measure, Pearson correlation coefficient was calculated between the estimated and original BDI values to determine how proportional the estimator is to the original BDI scale [26].

4 Results

4.1 Results of single channel model with MFCCs

The results obtained using 14 MFCC values are shown in Fig. 5. On the left vertical axis, the progress of the RMSE values on the test sets can be seen as a function of the epoch. The right vertical axis shows the training loss.

For both offsets, there is a monotonous decrease in RMSE, which is similar to that of training loss. It can be seen that applying 10 offset (submatrix's size 10 × 10) in

the correlation structure, lower RMSE was achieved than using 20 offset. At the end (500 epoch) of the training, 11.69 RMSE was resulted with $k = 20$, while 11.04 with $k = 10$.

The use of a large offset number is likely to make the correlation structure even more complex, which may carry additional information to estimate severity. However, it might require a deeper network, for which there is not enough data currently.

Using half of the MFCC features, the results are shown in Fig. 6. The axes and their names are the same as in Fig. 5. Applying a k of 10 gave an increase in RMSE after a given number of epoch. In this case, the network may not have found a pattern that it could use to estimate BDI effectively.

Similarly to Fig. 5, a decaying curve can be observed using $k = 20$ on Fig. 6. In this case, the more complex spectral structure contributed to the stable descending nature of the curve. Moreover, it has a lower RMSE than applying 10 offset after 370 epochs.

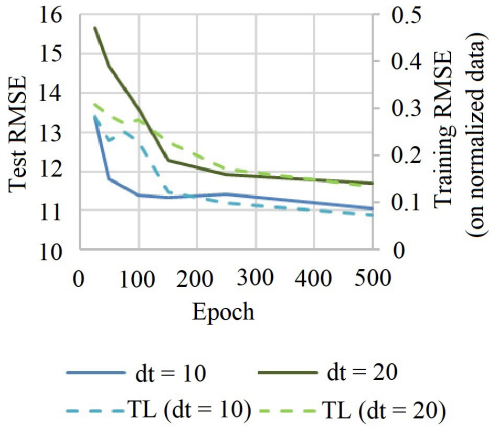

Fig. 5 Results obtained with 14 MFCC features on the single channel model with two different offset values. TL = training loss.

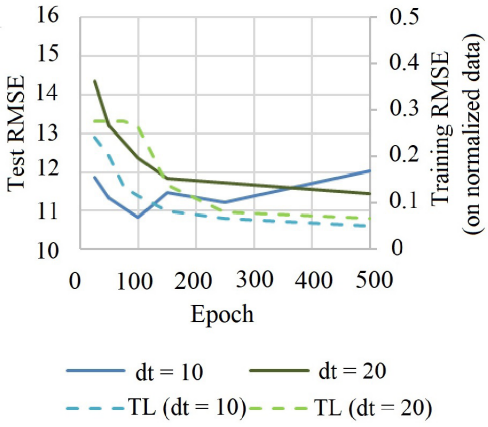

Fig. 6 Results obtained with 7 MFCC features on the single channel model with two different offset values. TL = training loss.

The training loss decreases earlier and reaches almost a constant value rather than in the Fig. 5. Thus, the test sample as able to follow the RMSE evolution of the training sample in a monotone manner at 20 offset. Increasing the epoch would probably have resulted lower RMSE using $k = 20$ (Fig. 6) or 14 MFCCs (Fig. 5).

4.2 Results of double channel model

In the single channel model, it was evident that using 10 offset, a lower RMSE was obtained with a given set of features in most cases than using 20 offset. Therefore, in the double channel model, MFCCs parallel with formant frequencies were investigated using only 10 offset.

The result is shown in Fig. 7, where the horizontal axis is the number of epoch. The left vertical axis includes the RMSE values of the test set while the right vertical axis consists of RMSE values of the training set. Both 7 and 14 MFCCs were examined separately with the specified formant frequencies as below:

Setup 1: 7 MFCCs + formants (referred as MFCC7),

Setup 2: 14 MFCCs + formants (referred as MFCC14).

It can be seen that the MFCC7 curves follow nearly the same progress as the 10 offset experiment in Fig. 6. The MFCC7 has a local minimum of test's RMSE at 150 epochs, then diverges at larger epochs. The training set RMSE decreases monotonously and converges to 0.

The MFCC14 also follows a decreasing progress similar to the 10 offset experiment in the Fig. 5. At lower number of epoch, it has spikes on the diagram. At higher epoch numbers, however, it takes on a monotonous decreasing characteristic. Based on the decline trend, lower RMSEs may be achievable with higher epoch numbers.

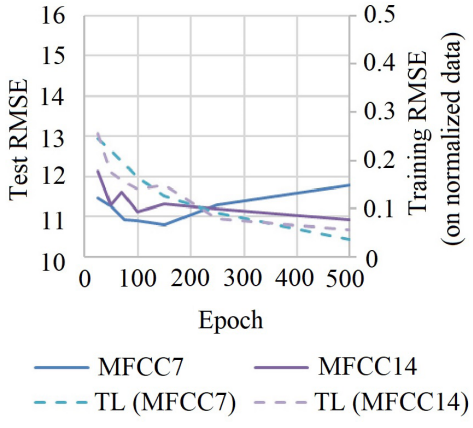

Fig. 7 Results achieved with MFCCs and the three formants with 10 offset. TL = Training loss.

The network also started at a lower RMSE, and was able to go below a values of 11 at certain intervals.

It should be taken into account that despite of the full cross-validation and regulation, over-learning is still possible. Therefore, an epoch can be selected where the training and testing loss the closest to each other. In other words, both training and testing curves have declined until the selected epoch, but after that, the curve of the test set may diverge. Table 2 summarizes the test RMSE and correlation at selected epochs.

It should also be borne in mind that if a point is selected too early as a local minimum, then the system may be undertrained at that point.

It can be seen that the test RMSE ranges from 11.7 to 10.8. Furthermore, the correlation values are at the top of the moderate and the bottom of the strong correlation category [27]. Based on this selection, the double input model performed better than the single ones.

The original and estimated BDI scatter plot is visualized on Fig. 8 at the correlation result of 0.61. It can be seen that it has a high standard deviation at higher BDI values.

Furthermore, RMSE values at the 0.61 correlation setting were derived according to the severity classes of depression (defined in Table 1). It can be found that our algorithm is more prone to error in the severe category than in the normal, mild and moderate cases (Table 3). According to this, the learning algorithm estimated depression of normal, mild, and moderate severity with an average 8.67 RMSE while the severe category rose up to 17.35 RMSE.

Referring to Table 3, acoustic features extracted from speech may have a less significant effect on the separation of depression at a severe level. Also, there are few samples that have outstandingly high BDI-II value in the database, making the CNN difficult to learn and predict more precisely in this category.

Projecting the result of MFCC7 and formants double channel model (at 150 epoch) to binary classification, the confusion matrix according to Table 4 can be written. 74 true positive, 81 true negative, 17 false negative and 10 false positive samples can be seen resulting in 85.2 % accuracy, 81.3 % sensitivity and 89.0 % specificity measures. Thus, a hypothetical ideal model would use the double channel approach with the given parameters of Section 3.4 and the evaluation of Section 3.5. The input features are MFCCs (displacement rate of 8 and 10 offset) and formants (displacement rate of 1 and 10 offset) that resulted less than 11 RMSE and strong correlation (0.60 and 0.61).

Table 2 Selected epochs where possible to stop the process.

k	Features	Epoch	RMSE test	*r*
10	14 MFCCs	500	11.045	0.60
20		500	11.686	0.54
10	7 MFCCs	250	11.210	0.58
20		500	11.433	0.55
10	7 MFCCs + 3 formants	150	10.797	0.61
	14 MFCCs + 3 formants	500	10.927	0.60

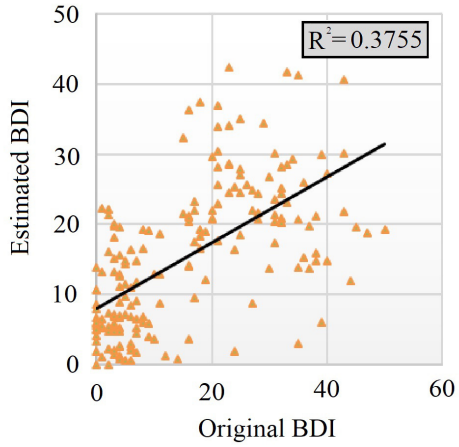

Fig. 8 Original and estimated BDI values for the double channel model with MFCC7 and formant frequencies (offset 10).

5 Conclusion

Acoustic features were obtained from speech recordings in the present study from which a special auto- and cross correlation structure were developed. The first 14 and 7.

MFCC values and the first three formant frequencies were used as features to estimate BDI values.

As regressor, convolutional neural network was used fed by images as input. Two models were created: single channel (MFCCs) and double channel (MFCCs and formant frequencies) models. The models were trained and

Table 3 Distribution of RMSE by depression severity classes based on Table 1 with MFCC7 and formants double channel model.

	Normal	Mild	Moderate	Severe
RMSE	7.58	9.56	8.88	17.35

Table 4 Result of binary classification with MFCC7 and formants features on the double channel model (at 150 epoch).

		Original	
		DE	**HC**
Predicted	**DE**	74	10
	HC	17	81

tested by full cross-validation and ADAM optimization.

Using the single channel model, the first 14 MFCCs achieved more monotonic decrease and almost resulted the same RMSE values than the 7 MFCCs.

The number of offsets (k) also affected the results. We found that increasing the offset number may provide more information in the correlation table. This is shown in the monotonous progress of the curves in case of MFCC7. However, the estimation in this case has worse (higher) RMSE statistics.

Using the dual channel model, better results were achieved at the certain epochs.

In terms of estimating the severity of depression, the RMSE had a flat tail after reaching a certain epoch number. Therefore, the experiment was stopped at 500 epochs as a compromise in running time and performance.

The obtained results are difficult to compare to others' works due to the distinct dataset (different in language, BDI distribution, sample number). We could not exceed the estimation of depression severity presented in Williamson's study (7.42 RMSE) [18]. However, it performed better than the specified baseline value (RMSE = 4.120) and was near to Lang He study (RMSE = 10.456) [19]. The comparisons can hardly be interpreted. For example, at high BDI values, the presented method resulted much higher RMSE values. We can easily imagine that different BDI distribution of a dataset results in higher or lower RMSE variations.

The advantage of the procedure is that it is independent of the subject's gender and does not require complicated preprocessing of the speech sample (e.g., segmentation at the level of speech). Finally, it should be noted that the regulations have been designed to minimize the possibility of over-learning.

This study has highlighted the possibility of using CNN to estimate the severity of depression. We definitely want to continue our research in the future. Possible directions are to include other features like prosodic ones, restructure the CNN architecture (changing layers and parameters) and data augmentation. The correlation structure can also be modified to better suit deep learning.

Acknowledgement

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.

References

[1] James, S., Abate, D., Abate, K., Abay, S., Abbafati, C., Abbasi, N., Murray, C. "Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990-2017: a systematic analysis for the Global Burden of Disease Study 2017", The Lancet, 392(10159), pp. 1789–1858, 2018. https://doi.org/10.1016/S0140-6736(18)32279-7

[2] Friedrich, M. J. "Depression is the leading cause of disability around the world", Jama Network, 317(15), pp. 1517–1517, 2017. https://doi.org/10.1001/jama.2017.3826

[3] Deckersbach, T., Dougherty, D. D., Rauch, S. L. "Functional imaging of mood and anxiety disorders", Journal of Neuroimaging, 16(1), pp. 1–10, 2006. https://doi.org/10.1177/1051228405001474

[4] Olesen, J., Gustavsson, A., Svensson, M., Wittchen, H. U., Jönsson, B. "The economic cost of brain disorders in Europe", European Journal of Neurology, 19(1), pp. 155–162, 2012. https://doi.org/10.1111/j.1468-1331.2011.03590.x

[5] Brådvik, L. "Suicide Risk and Mental Disorders", International journal of environmental research and public health, 15(9), Article number: 2028, 2018. https://doi.org/10.3390/ijerph15092028

[6] Orozco, J. R., Hoenig, F., Arias-Londoño, J. D., Vargas-Bonilla, J. F., Daqrouq, K., Skodda, S., Rusz, J., Nöth, E. "Automatic detection of Parkinson's disease in running speech spoken in three different languages", The Journal of the Acoustical Society of America, 139(1), pp. 481–500, 2016. https://doi.org/10.1121/1.4939739

[7] Tulics, M. G., Vicsi, K. "The automatic assessment of the severity of dysphonia", International Journal of Speech Technology, 22(2), pp. 341–350, 2019. https://doi.org/10.1007/s10772-019-09592-y

[8] Sztahó, D., Kiss, G., Tulics, M. G., Dér-Hajduska, B., Vicsi, K. "Automatic discrimination of several types of speech pathologies", In: 10th Conference on Speech Technology and Human-Computer Dialogue (SpeD 2019), Timisoara, Romania, 2019, pp. 1–6. https://doi.org/10.1109/SPED.2019.8906556

[9] Sztahó, D., Kiss, G., Tulics, M. G., Vicsi, K. "Automatic Separation of Various Disease Types by Correlation Structure of Time Shifted Speech Features", In: 41st International Conference on Telecommunications and Signal Processing (TSP), Athens, Greece, 2018, pp. 1–4. https://doi.org/10.1109/TSP.2018.8441395

[10] Cummins, N., Scherer, S., Krajewski, J., Schnieder, S., Epps, J., Quatieri, T. F. "A review of depression and suicide risk assessment using speech analysis", Speech Communication, 71, pp. 10–49, 2015. https://doi.org/10.1016/j.specom.2015.03.004

[11] Low L.-S., Maddage, M., Lech, M., Sheeber, L., Allen, N. "Influence of acoustic low-level descriptors in the detection of clinical depression in adolescents", In: 2010 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP), Dallas, TX, USA, 2010, pp. 5154–5157. https://doi.org/10.1109/ICASSP.2010.5495018

[12] Valstar, M., Schuller, B., Smith, K., Eyben, F., Jiang, B., Bilakhia, S., Schnieder, S., Cowie, R., Pantic, M. "Avec 2013: the continuous audio/visual emotion and depression recognition challenge", In: Proceedings of the 3rd ACM International Workshop on Audio/Visual Emotion Challenge (ACM Multimedia Conference), Barcelona, Spain, 2013, pp. 3–10. https://doi.org/10.1145/2512530.2512533

[13] Yang, Y., Fairbairn, C., Cohn, J. F. "Detecting depression severity from vocal prosody", IEEE Transactions Affective Computing, 4(2), pp. 142–150. 2013. https://doi.org/10.1109/T-AFFC.2012.38

[14] Zhou, Y., Sun, Y., Zhang, J., Yan, Y. "Speech Emotion Recognition Using Both Spectral and Prosodic Features", In: 2009 International Conference on Information Engineering and Computer Science, Wuhan, China, 2009, pp. 1–4. https://doi.org/10.1109/ICIECS.2009.5362730

[15] Cannizaro, M., Harel, B., Reilly, N., Chappell, P., Snyder, J. P. "Voice acoustical measurement of the severity of major depression", Brain and Cognition, 56(1), pp. 30–35. 2004. https://doi.org/10.1016/j.bandc.2004.05.003

[16] Hardy, P., Jouvent, R., Widlocher, D. "Speech pause time and the retardation rating scale for depression (ERD). Towards a reciprocal validation", Journal of Affective Disorders, 6(1), pp. 123–127. 1984. https://doi.org/10.1016/0165-0327(84)90014-4

[17] Mundt, J. C., Snyder, P. J., Cannizzaro, M. S., Chappie, K., Geralts, S. D. "Voice acoustic measures of depression severity and treatment response collected via interactive voice response (IVR) technology", Journal of Neurolinguistics, 20(1), pp. 50–64. 2007. https://doi.org/10.1016/j.jneuroling.2006.04.001

[18] Williamson, J. R., Quatieri, T. F., Helfer, B. S., Horwitz, R., Yu, B., Mehta, D. D. "Vocal biomarkers of depression based on motor incoordination", In: Proceedings of the 3rd ACM International Workshop on Audio/Visual Emotion Challenge (ACM Multimedia Conference), Barcelona, Spain, 2013, pp. 41–48. https://doi.org/10.1145/2512530.2512531

[19] Lang, H., Cui, C. "Automated depression analysis using convolutional neural networks from speech", Journal of Biomedical Informatics, 83, pp. 103–111, 2018. https://doi.org/10.1016/j.jbi.2018.05.007

[20] Smarr, K. L., Keefer, A. L. "Measures of depression and depressive symptoms: beck depression inventory-II (BDI-II), Center for Epidemiologic Studies Depression Scale (CES-D), geriatric depression scale (GDS), hospital anxiety and depression scale (HADS), and patient health Questionnaire-9 (PHQ-9)", Arthritis care & research, 63(S11), pp. S454-S466, 2011. https://doi.org/10.1002/acr.20556

[21] Beck, A. T., Steer, R. A. Garbin, M. G. "Psychometric properties of the Beck Depression Inventory: twenty-five years of evaluation", Clinical Psychology Review, 8(1), pp. 77–100, 1988. https://doi.org/10.1016/0272-7358(88)90050-5

[22] Boersma, P., Weenink, D. "PRAAT, a system for doing phonetics by computer", Glot International. 5(9/10). pp. 341–345, 2001. [online] Available at: https://www.fon.hum.uva.nl/paul/papers/speakUnspeakPraat_glot2001.pdf [Accessed: 20 February 2020]

[23] Kingma, D. P., Ba, J. "Adam: A method for stochastic optimization", arXiv preprint arXiv:1412.6980, 2017. [online] Available at: https://arxiv.org/pdf/1412.6980v9.pdf [Accessed: 25 February 2020]

[24] Jenei, A. Z., Kiss, G. "Possibilities of Recognizing Depression with Convolutional Networks Applied in Correlation Structure", In: 43rd International Conference on Telecommunications and Signal Processing (TSP), Milan, Italy, 2020, pp. 101–104. https://doi.org/10.1109/TSP49548.2020.9163547

[25] Chai, T., Draxler, R. R. "Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature", Geoscientific Model Development, 7(3), pp. 1247–1250, 2014. https://doi.org/10.5194/gmd-7-1247-2014

[26] Mukaka, M. M. "Statistics corner: A guide to appropriate use of correlation coefficient in medical research", Malawi Medical Journal, 24(3), 69–71, 2012. [online] Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3576830/ [Accessed: 25 February 2020]

[27] Norzehan, S., Noor, E. A. K., Inda, I. N. A. "Structuring elements of hit or miss to identify pattern of benchmark Latin alphabets strokes", Indonesian Journal of Electrical Engineering and Computer Science, 12(1), pp. 356–362, 2018. https://doi.org/10.11591/ijeecs.v12.i1.pp356-362