AUTOMATIC ASSESSMENT OF THE DEGREE OF CLINICAL DEPRESSION FROM SPEECH USING X-VECTORS

José Vicente Egas-López[1], *Gábor Kiss*[2], *Dávid Sztahó*[2], *Gábor Gosztolya*[1,3]

[1] University of Szeged, Institute of Informatics, Szeged, Hungary
[2]Budapest University of Technology and Economics, Budapest, Hungary
[3]MTA-SZTE Research Group on Artificial Intelligence, ELRN, Szeged, Hungary

ABSTRACT

Depression is a frequent and curable psychiatric disorder, detrimentally affecting daily activities, harming both workplace productivity and personal relationships. Among many other symptoms, depression is associated with disordered speech production, which might permit its automatic screening by means of the speech of the subject. However, the choice of actual features extracted from the recordings is not trivial. In this study, we employ x-vectors, a DNN-based feature extractor technique, to detect depression from a Hungarian corpus. We experiment with training custom x-vector extractors, and we also explore the performance of an *out-of-domain* pre-trained one. Our findings confirm that x-vectors are able to capture meaningful speaker traits that contain information for depression discrimination. We also show that the language of the extractor is of secondary importance compared to the frame-level feature set: our best model, which achieved an AUC score of 0.940 and an RMSE score of 9.54, was trained on log-energies instead of MFCCs.

Index Terms— speech processing, depression screening, x-vectors, pre-trained, i-vectors

1. INTRODUCTION

The speech is a biomarker containing information about a wide variety of traits (e.g., the mental status of the speaker). Depression is a psychiatric disorder affecting the patient on a wide scale. Although it is a frequent and curable disease, estimating its occurrence is hard due to the specific clinical expertise needed [1]. The fact that there may be a connection between depression and speech was pointed out by Kraepelin [2], one of the founders of modern psychology. Early examinations dealt with the analysis of individual speech features, and reported the decrease in the mean and dynamics of pitch values, slower articulation tempo [3] along with monotonous and lifeless dynamics [4, 5].

Various studies have investigated the possibility of assessing depression from speech. For instance, using CNNs for the enhancement of the detection of depression [6]; the analysis of gender and identity issues from the patients [7]; or feature extraction from the motor incoordination [8]. Here, we present an approach based on DNN embeddings (i.e., x-vectors [9]), to assess depression using the Hungarian Depressed Speech Dataset (HDSDb). This technique employs a DNN to map variable-length utterances to fixed-dimensional embeddings that contain meaningful speaker traits (e.g., the speaking style or the emotion [10]) which can be adapted to depression detection. Prior studies made use of the HDSDb, but with fewer samples, e.g.: CNNs and a speech correlation structure were used in [11] (accuracy of 84.1% with 188 samples). Also, the use of a special feature acoustic parameter selection approach in [12] (8.10 of RMSE with 127 samples). Both studies relied on Leave-One-Out Cross-Validation (LOOCV).

Here, we explore the sufficiency of x-vectors as a more straightforward method for discriminating the degrees of depression (i.e., the Beck Depression Inventory (BDI) II scale). Our evaluation is based on a, more impartial, speaker-wise Nested Cross-Validation instead of LOOCV. Also, we evaluate the model from a two-class perspective as we turn the predictions into binary labels based on a specific BDI threshold value for depression. Furthermore, we carry out an automatic feature selection method based on the Pearson's correlation of the features with respect to the BDI labels. Our key contributions are: (I) training custom x-vector models using language-domain matching, and exploring with data augmentation for training the extractors; (II) investigating the performances of the x-vector architecture on log-energies, and analyzing their effectiveness over cepstrum-based; (III) experimenting with the robustness of the embeddings (from custom and pre-trained models) after performing feature selection for this particular dataset. To the best of our knowledge, no other studies have so far adapted the x-vector technique for

This research was supported by grant NKFIH-1279-2/2020 of the Ministry for Innovation and Technology, Hungary, and by the Ministry of Innovation and Technology NRDI Office by grant no. NKFIH FK-124413, and within the framework of the Artificial Intelligence National Laboratory Program (MILAB). G. Gosztolya was also funded by the János Bolyai Scholarship of the Hungarian Academy of Sciences, and by the Ministry of Innovation and Technology New National Excellence Program ÚNKP-21-5-SZTE.

the screening of the levels of clinical depression.

2. DATA

We relied on the Hungarian Depressed Speech Dataset (HDSDb) [13]. The degree of severity of depression was recorded using the Beck Depression Inventory II (BDI) scale [14]. The BDI-II scale ranges from 0 (healthy state), to 63 (severe condition). This scale uses the following rating: 0-13 healthy, 14-19 mild depression, 20-28 moderate depression, 29-63 severe depression. The corpus consists of 222 speakers, 116 patients suffering from depression (mean BDI for males and females are 24.9 ($\pm$7.4) and 27.8 ($\pm$9.2), respectively) and 106 healthy control speakers (mean BDI for males and females are 4.3 ($\pm$3.4) and 4.2 ($\pm$3.0), respectively) with a balanced age distribution and 146 males. The speakers read a short tale ('The North Wind and the Sun'). The recordings were sampled at 16 kHz and 16-bit.

3. DNN EMBEDDINGS: X-VECTORS

The x-vector approach can be thought as of a neural network feature extraction technique that provides fixed-dimensional embeddings corresponding to variable-length utterances. The architecture of the DNN is as follows: the *frame-level* layers have a time-delay architecture, and let us assume that t is the actual time step. At the input, the frames are spliced together; namely, the input to the current layer is the spliced output of the previous layer. Next, the *stats pooling* layer gets the T frame-level output from the last frame-level layer (*frame5*), aggregates over the input segment, and computes the mean and standard deviation. The mean and the standard deviation are concatenated and used as input for the next *segment* layers; from any of these layers the *x-vectors* embeddings can be extracted. And finally, the *softmax* output layer (which is discarded after training the DNN) [9, 15, 16]. Instead of predicting frames, the network is trained to predict speakers from variable-length utterances.

The embeddings produced by the network described above capture information from the speakers over the whole audio-signal. These are the *x-vectors* and can be extracted from any *segment* layer. We rely on the x-vector approach since it acquires meaningful characteristics (i.e., speaking-style information and emotion [10]) at utterance level rather than at frame level, which results in a fixed-sized vector irrespective of the length of the utterance.

4. EXPERIMENTS AND RESULTS

4.1. DNN extractor training

Training neural networks generally implies having a significant amount of samples for getting a good performance; however, the HDSDb is quite limited. A DNN that learns from this kind of data would result in under-fitting; and consequently would over-fit the final classifier. Hence, we did not make use of the HDSDb corpus to train any extractor. We fitted two different x-vector extractors using distinct corpora: *first*, we employed a subset of 60 hours (10,636 utterances) of the *BEA Corpus* [17] (Hungarian spontaneous speech). And *second*, we utilized the pre-trained x-vector model [9] that was fitted on English speech corpora (Switchboard (SWBD) plus NIST SRE). Besides investigating the usefulness of the pre-trained model on a different type of task, we also sought to discover the difference in quality of x-vector representations extracted using distinct models, which differ in both amount and language in terms of their training data.

We used two types of frame-level representations: 23 Mel-Frequency Cepstral Coefficients (MFCC) and 40 filterbanks (FBANKs); both with a frame-length of 25ms, and a frame-shift of 10ms. While MFCCs are the standard for fitting x-vector models, FBANKs have proved to be effective in deep learning studies related to speech analysis, e.g., in speech recognition tasks [18, 19]. Furthermore, in a previous work [20], we also demonstrated the usefulness of applying log-energies over MFCCs in x-vector training.

4.2. BEA Corpus Augmentation

Seeking to improve the diversity of the data and the noise robustness of the model, we carried out data augmentation on the BEA corpus. The augmented dataset was used to fit two additional extractors. The augmented versions were added by choosing randomly from the following types: babble, music, noise, and reverberation. The first three correspond to adding or fitting noise to the original utterances. The fourth one involves a convolution of room impulse responses with the audio. The augmentation procedure increased the BEA corpus to 52,636 utterances (293 hours). We sought to evaluate the contribution of the augmentation methods to the quality of the embeddings. Nevertheless, as with previous findings (see [21]), adding noise and reverberation for x-vector training does not always lead to robustness and might be dependent on the quality of the utterances.

4.3. Baseline Approach

We opted for a former state-of-the-art speaker recognition method: the i-vector approach, which is known to capture speech, speaker and utterance meta information [22]. Akin to x-vectors, i-vectors also contain relevant information within the channel factor, which was used to classify emotion before [23]. Moreover, i-vectors have been successfully adapted to depression screening tasks giving good performances [24, 25]. Here, we trained the GMM-UBM model utilizing the same corpus (i.e., the BEA) that was employed for training the *first* x-vector extractor. The GMM-UBM was fitted with 256 Gaussian components, which was used to extract i-vector representations from the HDSDb dataset.

Table 1. Results of the experiments using all the feature dimensions.

	Regression		Classification				
	Pearson's CC	**RMSE**	**UAR**	**SPEC**	**SENS**	**AUC**	**F1**
i-vector baseline	.608	10.45	80.44	89.39	82.00	0.920	89.36
BEA Extractor (FBANK)	.625	10.26	88.65	86.79	90.51	0.920	89.36
BEA Extractor (MFCC)	.615	10.36	82.64	77.35	87.93	0.904	84.29
BEA-augmented Extractor (FBANK)	**.684**	**9.54**	**89.00**	**84.90**	**93.10**	**0.940**	**90.00**
BEA-augmented Extractor (MFCC)	.635	10.16	80.75	73.58	87.93	0.908	82.92
Pre-trained Model (MFCC) [9]	.675	9.64	82.99	75.47	90.51	0.935	85.02

4.4. Evaluation

Here, we made use of the Support Vector Regression algorithm (nu-SVR with linear kernel). The x-vector representations, after standardization, were fed to the regressor for training. In contrast to former studies on the same corpus, and, seeking to avoid an optimistically-biased evaluation of the model, we chose speaker-wise Nested Cross-Validation. The metrics employed were the Pearson's CC of the ground truth and predicted BDI scores of the subjects, along with the Root Mean Square Error (RMSE).

Besides the above, we also performed a classification performance analysis by transforming the labels into two classes. Namely, we evaluated our models from a binary class problem perspective. Thus, the subjects were automatically categorized as having depression or not by binarizing the labels based on their BDI values, where: if $BDI \geq 13.5$, the patient was cataloged as depressed; healthy control otherwise. This way, the class distribution resulted in 116 patients and 106 healthy controls. In this context, we selected various metrics that provided a broader picture of the performance of the transformed predictions. As in most medical research, we used sensitivity and specificity, F1-score, along with Unweighted Average Recall (UAR, being the mean of specificity and sensitivity), and Area Under the Receiver Operating Characteristics Curve (AUC).

4.5. Results and Discussion

Table 1 presents the results of our experiments along with the i-vector baseline, which was surpassed by the methods based on x-vectors. In general, the DNN embeddings could model better speaker traits for depression screening than the i-vectors. The augmented extractor produced better embeddings than their non-augmented counterparts, and demonstrated the effectiveness of data-augmentation when using x-vectors on this specific corpus. In particular, the extractor trained with the BEA-augmented corpus (with FBANKs) gave the highest Pearson's CC: .684, and the lowest RMSE: 9.54. As for the *binary* classification evaluation, the same

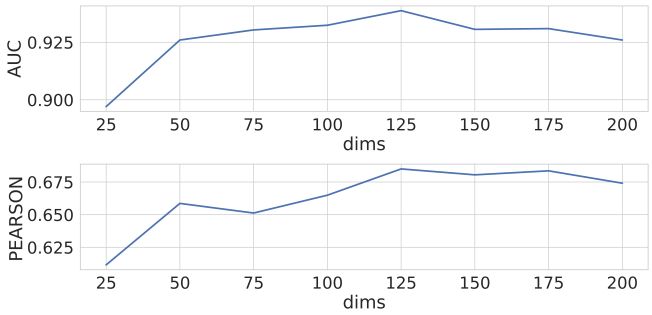

Fig. 1. CC and AUC scores of the feature selection process from the **BEA-augmented Extractor (FBANK)**.

configuration gave the highest scores: a UAR of 89.00%, an specificity of 84.90%, a sensitivity of 93.10%, a AUC-score of 0.940, and an F1-score of 90. We found a quite low number of false negative and false positive cases, which indicates the potential feasibility of the model for screening. Moreover, the AUC-score value suggests a considerably high discriminating ability of the model.

The extractors fitted with log-energies (except for the non-augmented version) outperformed their cepstra counterparts in every case. This may be due to the fact that MFCCs attempt to eliminate unimportant variations for recognition, and lead to a reduction in the input-signal dimension (less information). Meanwhile, the FBANKs contain a more integral representation as they produce a less pre-processed input-signal with a larger set of filter-bank coefficients (more information); it appears that DNNs are able to better exploit these type of representations. The embeddings from the *pre-trained model (MFCCs)*, however, achieved better scores than the *BEA Extractor (FBANKs)* configuration. Although our custom extractors used in-domain language data, a possible reason could be the huge difference between their corresponding amounts of training corpora.

The *Pre-trained Model [9]*, although competitive, could not surpass the results of the best custom model, we got a lower CC (.675). The results may confirm an existing *data-*

Table 2. Results of the experiments using the **correlation-based feature selection**. The best feature selection configurations are presented only. N denotes the number of features from the automatic feature selection process.

	Regression		**Classification**					
	Pearson's CC	**RMSE**	**UAR**	**SPEC**	**SENS**	**AUC**	**F1**	**N**
BEA Extractor (FBANK)	.632	10.14	87.19	83.01	91.34	0.915	88.33	150
BEA Extractor (MFCC)	.586	10.59	78.39	68.86	87.93	0.891	81.27	175
BEA-augmented Extractor (FBANK)	**.685**	**9.50**	**88.61**	**85.84**	**91.37**	**0.938**	**89.45**	125
BEA-augmented Extractor (MFCC)	.603	10.41	81.11	71.69	90.51	0.914	83.66	175
Pre-trained Model (MFCC) [9]	.672	9.70	83.89	76.41	**91.37**	0.934	85.82	200

domain dependence of the x-vector architecture. More specifically, we experimented with models that learned from data closer to the actual task domain (language-related in this specific case), and they produced better quality representations than the pre-trained model did.

5. CORRELATION-BASED FEATURE SELECTION

In fact, the x-vector features will have a bigger number of dimensions than the total number of samples of the dataset. Eventually, this might lead to a decay in the performance due to the regularization bias growth towards the training data. Hence, before training, we carried out an automatic feature selection, seeking to reduce the number of features. More precisely, we computed the CC for each feature-column with respect to the BDI labels; from these values, we selected those that had the highest CC scores. The final selection of the number of dimensions (N) was based on a step size of 25 (i.e., $N = 25, 50, \ldots, 200$ selected dimensions). The procedure was carried out within the speaker-wise Nested-CV to avoid peaking. Consequently, besides dimensionality reduction, it also meant that we just had relevant features (those that contribute the most to the final predictions), and thus speeded up the BDI estimation step.

The results of this approach are given in Table 2. Similar to the previous experiments, the augmented extractors fitted with FBANKs also outperformed the rest of the configurations in this case. Moreover, the CC increased slightly to .685, while the RMSE decreased to 9.50. These results were achieved just using 125 of the 512 available original features after the feature selection process. Also, the classification metrics for the same configuration changed slightly: while the specificity score experienced an increment, the sensitivity, AUC, UAR, and F1 scores only decreased a small amount. Again, FBANKs features gave more efficient performances based on the number of selected features. Fewer dimensions were needed for the model to provide a better generalization; that is, FBANK-based embeddings actually contained more meaningful information than those from MFCCs.

Figure 1 depicts the AUC and the Pearson's CC scores obtained using the different N feature selection values for the corresponding dimension size. The line plots display a tendency where the CC values increase as the number of dimensions increment as well, and they both start to decrease after dimension 150. In general, both metrics suggest quite similar trends over the number of dimensions. Overall, the correlation-based feature selection, besides discarding irrelevant information and helping to reduce the computation times, also helped to increase the CC and reduce the RMSE in most of the cases. Furthermore, all the configurations necessitated only less than the half of the original number of dimensions and produced better or competitive results.

6. CONCLUSIONS

This paper investigated the automatic estimation of the levels of clinical depression from the speech using speaker recognition methods. Specifically, we presented x-vector embeddings that contain information that is predictive of depression. Our custom x-vector extractors learned from distinct frame-level features acquired from corpora matching the language of the actual task. Also, we found an improvement of the quality of the embeddings when computing them using *augmented* x-vector models. In this context, we spotted a slight language-domain dependence of the x-vector method as our best tailored extractor surpassed the performance of the pre-trained model even after the feature selection process. Furthermore, our findings confirmed that log-energies appear to be a robust alternative of cepstra coefficients for x-vector training as they provide larger (and more informative) input representations. We showed how our correlation-based feature selection approach produced similar performance scores using only a quarter of the features. Finally, we presented highly competitive CC and RMSE scores compared to those from former studies that used the same corpus and based their evaluations using optimistic methods (i.e, LOOCV), which proves the effectiveness of our approaches.

7. REFERENCES

[1] Mary Jane Friedrich, "Depression is the leading cause of disability around the world," *Jama*, vol. 317, no. 15, pp. 1517–1517, 2017.

[2] Emil Kraepelin, "Manic depressive insanity and paranoia," *The Journal of Nervous and Mental Disease*, vol. 53, no. 4, pp. 350, 1921.

[3] Murray Alpert, Enrique R Pouget, and Raul R Silva, "Reflections of depression in acoustic measures of the patient's speech," *Journal of affective disorders*, vol. 66, no. 1, pp. 59–69, 2001.

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

[5] Daniel M Low, Kate H Bentley, and Satrajit S Ghosh, "Automated assessment of psychiatric disorders using speech: A systematic review," *Laryngoscope Investigative Otolaryngology*, vol. 5, no. 1, pp. 96–116, 2020.

[6] Z. Huang, J. Epps, D. Joachim, B. Stasak, J. Williamson, and T. Quatieri, "Domain adaptation for enhancing Speech-based depression detection in natural environmental conditions using dilated CNNs," *Proceedings of Interspeech 2020*, pp. 4561–4565, 2020.

[7] P. Lopez-Otero and L. Docio-Fernandez, "Analysis of gender and identity issues in depression detection on de-identified speech," *Computer Speech & Language*, vol. 65, pp. 101118, 2021.

[8] J. Williamson, T. Quatieri, B. Helfer, R. Horwitz, B. Yu, and D. Mehta, "Vocal biomarkers of depression based on motor incoordination," in *Proceedings of AVEC*, 2013, pp. 41–48.

[9] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker verification," in *Proceedings of ICASSP*, 2018, pp. 5329–5333.

[10] J. Williams and S. King, "Disentangling style factors from speaker representations.," in *Proceedings of Interspeech*, 2019, pp. 3945–3949.

[11] Attila Zoltán Jenei and Gábor Kiss, "Possibilities of recognizing depression with convolutional networks applied in correlation structure," in *2020 43rd International Conference on Telecommunications and Signal Processing (TSP)*. IEEE, 2020, pp. 101–104.

[12] G. Kiss and K. Vicsi, "Mono-and multi-lingual depression prediction based on speech processing," *International Journal of Speech Technology*, vol. 20, 2017.

[13] G. Kiss, M. Tulics, D. Sztahó, A. Esposito, and K. Vicsi, "Language independent detection possibilities of depression by speech," in *Recent advances in nonlinear speech processing*, pp. 103–114. Springer, 2016.

[14] JRZ Abela and DU D'Alessandro, "A test of the diathesis-stress and causal mediation components of beck's cognitive theory of depression," *British Journal of Clinical Psychology*, vol. 41, no. 1, pp. 1, 2002.

[15] D. Snyder, D. Garcia-Romero, D. Povey, and S. Khudanpur, "Deep Neural Network embeddings for text-independent speaker verification," in *Proceedings of Interspeech*, 2017.

[16] D. Snyder, P. Ghahremani, D. Povey, D. Garcia-Romero, Y. Carmiel, and S. Khudanpur, "Deep Neural Network-based speaker embeddings for end-to-end speaker verification," in *Proceedings of SLT*, 2016.

[17] T. Neuberger, D. Gyarmathy, T. E. Gráczi, V. Horváth, M. Gósy, and A. Beke, "Development of a large spontaneous speech database of agglutinative Hungarian language," in *Proceedings of TSD*, Brno, Czech Republic, Sep 2014, pp. 424–431.

[18] A. Mohamed, G. E Dahl, and G. Hinton, "Acoustic modeling using deep belief networks," *IEEE transactions on audio, speech, and language processing*, vol. 20, no. 1, pp. 14–22, 2011.

[19] H. Seki, K. Yamamoto, and S. Nakagawa, "A deep neural network integrated with filterbank learning for speech recognition," in *Proceedings of ICASSP*, 2017.

[20] J. V Egas-López, M. Vetráb, L. Tóth, and G. Gosztolya, "Identifying Conflict Escalation and Primates by Using Ensemble X-Vectors and Fisher Vector Features," *Proceedings of Interspeech*, pp. 476–480, 2021.

[21] J. V Egas-López and G. Gosztolya, "Deep Neural Network Embeddings for the Estimation of the Degree of Sleepiness," in *Proceedings of ICASSP*, 2021.

[22] N. Dehak, P. J Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end Factor Analysis for Speaker Verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[23] Igor Jauk, *Unsupervised learning for expressive speech synthesis*, Universitat Politècnica de Catalunya, 2017.

[24] A. Afshan, J. Guo, S. J. Park, V. Ravi, J. Flint, and A. Alwan, "Effectiveness of voice quality features in detecting depression," *Proceedings of Interspeech*, 2018.

[25] M. Senoussaoui, M. Sarria-Paja, J. F Santos, and T. H Falk, "Model fusion for multimodal depression classification and level detection," in *Proceedings of AVEC*, 2014, pp. 57–63.