AUTOMATIC DEPRESSION LEVEL ASSESSMENT FROM SPEECH BY LONG-TERM GLOBAL INFORMATION EMBEDDING

Ya Li[1]*, Mingyue Niu*[2]*, Ziping Zhao*[2]*, Jianhua Tao*[3]

[1]School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing, China
[2]School of computer and information engineering, Tianjin Normal University, Tianjin, China
[3]National Laboratory of Pattern Recognition, Institute of Automation,
Chinese Academy of Sciences, Beijing, China

yli01@bupt.edu.cn, mniu@tjnu.edu.cn, ztianjin@126.com, jhtao@nlpr.ia.ac.cn

ABSTRACT

Depression is a serious mood disorder which brings negative effects on people's social activities. Therefore, growing attention has been paid to automatic depression assessment, especially from speech. However, most of the previous work uses hand-crafted features or deep neural network-based feature extractors to obtain deep features and then feed them into a classifier or a regression, which ignores the temporal relation of these features. To address this issue, this paper proposes a global information embedding (GIE) to make use of the long-term global information of depression and re-weight the LSTM output sequence. The short-term features are then pooled into long-term features by LASSO optimization to further improve the accuracy of depression recognition. Experiments on AVEC 2013 and AVEC 2014 verified the proposed method, and the RMSEs are 9.63 and 9.40, respectively.

Index Terms— depression, global information embedding, LSTM, attention, speech processing

1. INTRODUCTION

As a mental disease, depression is a serious mood disorder, which could cause organic diseases and affect people's ability to participate in social activities and communication. The World Health Organization has estimated that by 2030, depression will become the first disease burden in the world. A lot of effort on speech-based machine learning method for depression assessment has already been made to alleviate the current medical condition, since previous work found that the voice of depressed patients is relatively low and untoned, and has less phonetic variety and low speed. Based on these qualitative results, most of the early work uses hand-craft features, e.g., format [1, 2], prosody [3, 4], spectrum [2, 5], voice quality [6, 7], and articulatory features [8] to automatic detect depression severity level from speech. In recent years, deep learning-based methods have improved the performance of many related tasks, including affective computing and depression recognition. Recurrent neural networks (RNNs), including its variant LSTM [9, 10], and convolutional neural networks (CNNs) [8, 11, 12, 13] are the two dominant methods used to improve the feature representation, compared with the conventional method. Alhanai et al. [9] modeled the audio and text features by LSTM for depression detection, and they achieved 0.77 in terms of F1 score on DAIC dataset. Niu et al. [10] proposed a hybrid network and used LSTM to extract the spatial and temporal changes of short-term MFCC segments at the same time. EmoAudioNet [14], which was an aggregation of the MFCC-based CNN and the spectrogram-based CNN was proposed to detect depression from speech, and the F1 score was 0.82 on DAIC-WOZ dataset.

The specific problem of depression speech is its emotional expression is weak and affected by a longer context [9], compared with the general emotion recognition problem. To this end, an I-vector based feature representation was proposed to convert the frame-level features to a global representation [15]. Seneviratne et al. [16] used a dilated CNN model to obtain segment-level predictions, and then used an RNN model to obtain session-level predictions from the segment-level predictions. The accuracies are 62.86% for three-class and 71%-82% for two-class on different datasets. Different pooling techniques in deep neural networks can also be considered as a way to model long-time information.

Although much progress in speech-based depression recognition has been made in recent years, there is still room for improvement, especially from the aspects of temporal modeling since this is an inherent characteristic of emotion recognition. Therefore, this paper proposes a global information embedding (GIE) method to extract the high-level information of long-term features and feed it into the modeling process of LSTM, to improve the long-term feature representation ability of the model. Afterward, the features generated by the GIE model are pooled into long-term features by LASSO optimization, instead of the traditional max or average pooling, to further improve the accuracy of depression recognition. The proposed method is tested on the commonly used datasets, i.e., AVEC 2013 [17] and AVEC 2014 [18], and the experimental results verify the effectiveness of the proposed method.

The rest of this paper is organized as follows. Section 2 presents the proposed global information embedding method and the whole architecture of the depression assessment system. Experimental results and discussion are shown in Section 3, and Section 4 concludes the paper.

2. PROPOSED METHOD

2.1. System description

The proposed system is based on the hybrid network architecture introduced in [10], which is shown by Figure 1. As MFCC is a discriminative biomarker for depression disorder detection [19], the system extracts short-term MFCC features of an utterance of speech first. The MFCC sequence is then divided into segments with a given length and used as the input of the hybrid network.

In the hybrid network, an LSTM and a CNN sub-networks are trained separately by minimizing the mean square error between

the sub-network outputs and Beck Depression Inventory-II (BDI-II) score [20]. The LSTM and CNN sub-networks are designed to capture the depression differences in temporal changes and spatial structure, respectively. To model the global temporal information for LSTM, the long-term global information embedding (GIE) is proposed in this paper to re-weight each frame in the segment. Meanwhile, a squeeze-and-excitation [21] based channel attention is applied to the CNN to refine the channel importance. These will be described in details in Sections 2.2 and 2.3, respectively.

After the sub-network is trained, hidden layer representations preceding the output layer of two sub-networks are concatenated together as segment-level features, and then for each utterance, all segment-level features are combined by ℓ_p norm pooling and LASSO to generate the utterance-level features. Finally, support vector regression is employed to predict the depression level for each utterance. More details of the hybrid network architecture can be found in [10].

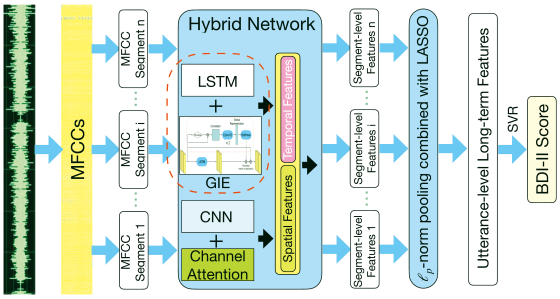


Fig. 1. The baseline hybrid network architecture.

2.2. LSTM with long-term global information embedding

The conventional LSTM leverages history information from the previous time instants to obtain the current representation. However, the symptoms of depression in speech may not be observed at early instants. This may lose the depression cues contained in the input sequence. To address the issue, this paper proposes a GIE module to make use of the long-term global information of depression.

The idea of GIE is similar to squeeze-and-excitation network, and is shown as (a) in Figure 2. The input feature sequence is denoted as $\boldsymbol{S} \in \mathbb{R}^{T\times D}$, where T and D are the segment length and input feature dimension of LSTM. The GIE first computes the correlation matrix of the input sequence, denoted by $\boldsymbol{M}$, to obtain second-order statistic information. This is written as

$$\boldsymbol{M} \quad = \quad \boldsymbol{S}\boldsymbol{S}^{\top} \in \mathbb{R}^{T\times T}. \tag{1}$$

Then, the correlation matrix is transformed by two 1-dimensional convolutional operations in series to generate the global representation. The kernel size of the first convolution is 1, which makes the convolutional operation become the multiplication between the input matrix and a vector $\boldsymbol{w}_1 \in \mathbb{R}^{T\times 1}$. The kernel size of the second convolution is 3, hence the convolutional operation can be represented as the multiplication between a block matrix $\boldsymbol{W}_2 \in \mathbb{R}^{T\times T}$ with block size 1×3 and the input vector. ReLU nonlinear activation function is applied after each convolution. Therefore, the global representation $\boldsymbol{g}$ can be computed by

$$\boldsymbol{g} \quad = \quad \delta(\boldsymbol{W}_2\delta(\boldsymbol{M}\boldsymbol{w}_1)) \in \mathbb{R}^{T\times 1} \tag{2}$$

where $\delta(\cdot)$ denotes the ReLU activation function. After generated, the global representation is embedded to the LSTM output by the attention mechanism, which is implemented as

$$\boldsymbol{S}_E \quad = \quad \text{Softmax}(\boldsymbol{g}) \circledast \text{LSTM}(\boldsymbol{S}) \in \mathbb{R}^{T\times D} \tag{3}$$

where $\boldsymbol{S}_E$ is the output embedding, and $\circledast$ denotes the extended matrix multiplication. The extended matrix multiplication between vector $\boldsymbol{x} \in \mathbb{R}^{T\times 1}$ and matrix $\boldsymbol{Y} \in \mathbb{R}^{T\times D}$ is defined as

$$\boldsymbol{x} \circledast \boldsymbol{Y} = [\boldsymbol{x}, ..., \boldsymbol{x}] \otimes \boldsymbol{Y} \tag{4}$$

where $\otimes$ denotes the element-wise multiplication. Thus, the attention mechanism is to re-weight the representation of LSTM output for each frame, such that the sequence focuses more on the frames with more useful depression cues. Finally, the LSTM input $\boldsymbol{S}$ is added to the output embedding $\boldsymbol{S}_E$ to compensate depression information from the original sequence.

The LSTM sub-network architecture is shown as (b) in Figure 2, which consists of N successive LSTM layers applying GIE. $N = 2$ is set in this paper. Then, a 1-dimensional convolutional layer with kernel size 1 is deployed following the final LSTM layer output. As mentioned above, the 1-dimensional convolution operation results in a T-dimensional vector, which is then transformed to a 64-dimensional temporal feature vector by a fully connected layer. The nonlinear activation functions in all layers are ReLU. The LSTM sub-network is trained to predict the BDI-II score. After trained, the temporal features will be used by the downstream operations.

2.3. CNN with channel attention

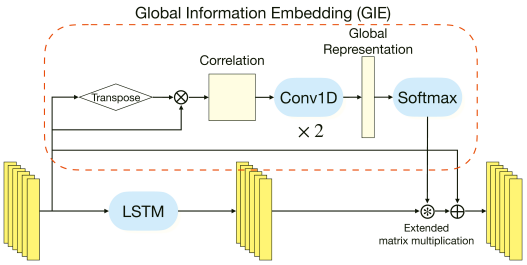
Apart from the temporal information captured by the LSTM sub-network, the spatial information is captured by the CNN sub-network. However, different CNN channels may have different abilities of depression cue mining. Therefore, it is important to highlight the channels which are more useful to depression assessment. For this purpose, a squeeze-and-excitation based channel attention shown as (c) in Figure 2 is applied to the CNN sub-network. In the squeeze part, a convolutional output tensor with size $T \times D \times R$ is transformed to an R-dimensional vector by a global average pooling, where R is the number of channels. In the excitation part, the R-dimensional vector is transformed by three 1-dimensional convolutional operations in series to generate the channel weights, where the kernel sizes are 3. Finally, each channel weight is added to the corresponding channel features of the original convolutional output tensor, to refine the importance of each channel.

The CNN sub-network architecture is shown as (d) in Figure 2, which consists of N successive 2-dimensional convolutional layers applying channel attentions, followed by successive 2-dimensional and 1-dimensional convolutional layers. $N = 2$ is set in this paper. Every 2-dimensional convolution has 64 kernels with size of 3×3. The 1-dimensional convolution has kernel size 1, and followed by a fully connected layer with 64 neurons to generate spatial features. The nonlinear activation functions in all layers are ReLU. The CNN sub-network is trained to predict the BDI-II score, and the spatial features will be used by the downstream operations being trained.

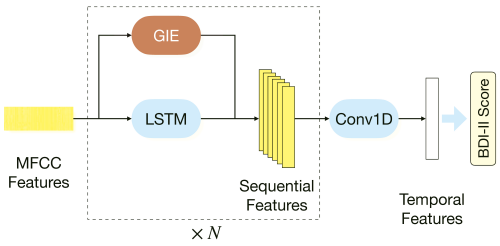
3. EXPERIMENTS

3.1. Data description and experimental setup

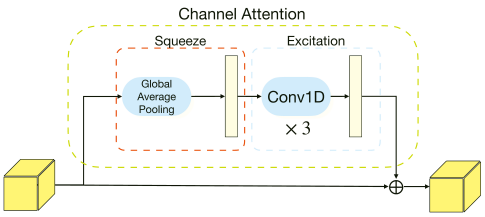
The AVEC 2013 [17] and AVEC 2014 [18] depression corpora were utilized in the experiments for evaluating the proposed method. The AVEC 2013 corpus contains 150 video samples of 14 tasks recorded from 84 subjects, of which the durations are between 20 to 50 minutes, and 25 minutes on average. Among these video samples, 77



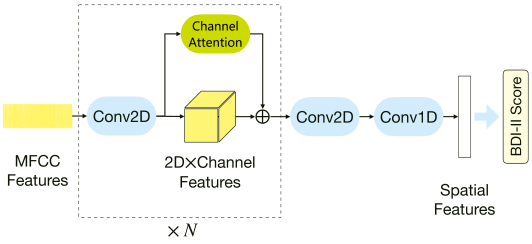
(a) The proposed long-term global information embedding.



(b) The LSTM sub-network architecture with GIE.



(c) The squeeze-and-excitation based channel attention.



(d) The CNN sub-network architecture with channel attention.

Fig. 2. The architectures of (a) proposed long-term global information embedding (GIE), (b) LSTM sub-network with GIE, (c) squeeze-and-excitation based channel attention, and (d) CNN sub-network with channel attention.

and 73 of them are corresponding to the healthy and depressed subjects, respectively. The dataset was divided into the training set, development set, and test set, of which each consisted of 50 video samples. The AVEC 2014 corpus is a subset of AVEC 2013, which only contains two tasks: Northwind and Freeform. Each task contains 150 video samples. The two tasks were combined in the experiments and equally divided into the training set, development set, and the test set, and thus there are 100 video samples for each set.

For evaluation and analysis, BDI-II scores ranging from 0 to 63 were employed as the assessment criteria to measure the depression severity. Four depression levels including no depression (0-13 scores), mild depression (14-19 scores), moderate depression (20-28), and severe depression (29-63 scores) were assigned to each sample. Root mean square error (RMSE) and mean absolute error (MAE) between the true and predicted BDI-II scores were exploited to evaluating the method performance. They are computed by

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}, \quad \text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \qquad (5)$$

where y_i and $\hat{y}_i$ are the true and predicted BDI-II scores of the ith sample among the N samples in total.

Only the speech was used in the proposed system, where 39 MFCC features were extracted using a window of 10 milliseconds with a shift of 5 milliseconds. The length of each segment in an utterance was 2.5 seconds, with 50% overlap on the adjacent segments. The networks were built as described in Section 2.

3.2. Results and discussion

3.2.1. Ablation tests and analysis

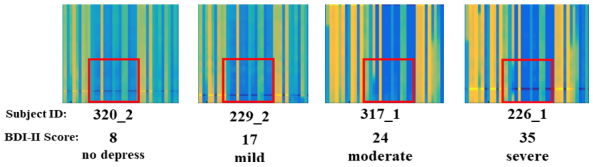
The ablation tests were conducted using the development sets to verify the effectiveness of the proposed method. The results are shown in Table 1. The network architecture of LSTM with GIE in the table is shown by (b) in Figure 2 and described in Section 2.2, while that of the comparable conventional LSTM network was configured by just removing the GIE block (the red block) in the figure. Table 1 shows that the proposed LSTM network with GIE consistently outperformed the conventional LSTM network in both RMSE and MAE, on both the AVEC 2013 and AVEC 2014 development sets. This suggests the effectiveness of correlation information in the proposed GIE, which leverages the global information of depression from the input sequence and complements the LSTM output sequence, whereas the conventional LSTM can only make simple use of the history information.

Table 1. The prediction performance of depression level using different network architectures on the development sets.

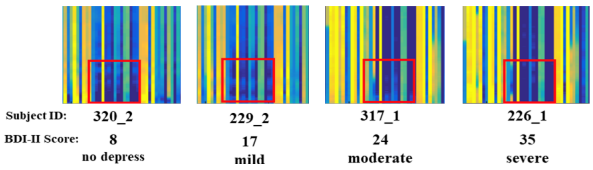
Network	AVEC 2013		AVEC 2014	
	RMSE	MAE	RMSE	MAE
Conventional LSTM	8.52	6.77	8.14	6.45
LSTM + GIE	8.27	6.31	8.03	6.12
Conventional CNN	8.82	6.59	8.53	6.76
CNN + Channel attention	8.56	6.23	8.33	6.47
Proposed Hybrid	**7.65**	**6.01**	**7.32**	**5.39**

To further analyze the effectiveness of GIE, the feature sequences in various depression levels output from the final LSTM layer with or without GIE were compared and shown visually in Figure 3. Speech utterances of four subjects with depression levels of no depression, mild depression, moderate depression, and severe depression were utilized for the analysis. In (a) of Figure 3 which is produced by conventional LSTM, the features framed in the red color show almost no difference between no depression and mild depression, as well as between moderate and severe depression. On the contrary, the features framed in red color generated by LSTM with GIE, which is shown by (b) of the figure, present significant differences between each pair of the four depression levels. This

suggests that the proposed GIE is helpful to improve the LSTM for capturing depression information.



(a) The LSTM output in 4 different depression levels without using GIE.



(b) The LSTM output in 4 different depression levels using GIE.

Fig. 3. Visual results for temporal features of individuals with various depression levels.

On the other hand, different channels in the convolutional layer may capture different local cues of depression for assessment. Channel attention makes the downstream convolution focus more on the input channels with more useful cues. The CNN with channel attention in Table 1 was built as (d) in Figure 2 and Section 2.3, while the comparable conventional CNN network was built by removing the channel attention block (the green block) in the figure. As expected, the comparison in Table 1 shows that the use of channel attention could improve the CNN model consistently.

Figure 4 visually shows the features of the 8th, 16th, and 32th channels in various depression levels and the output from the final 2-dimensional convolution. It is obvious that different channels have different abilities of mining depression cues. This may explain the effectiveness of the channel attention for depression detection.

Finally, the proposed hybrid network in Table 1 combined the sub-networks of LSTM with GIE and CNN with channel attention, which is shown by Figure 1 and described in Section 2.1. It significantly outperformed the above sub-networks on both AVEC 2013 and AVEC 2014 development sets by up to 1.0 in RMSE.

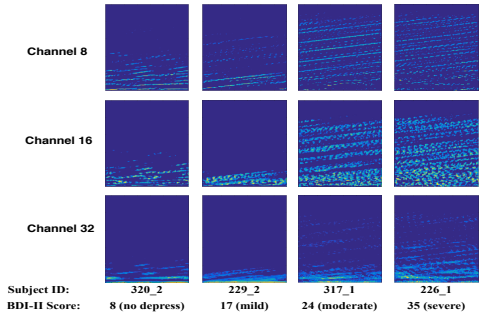


Fig. 4. Visual results for spatial features of individuals with various depression levels.

3.2.2. System comparison

The test sets of AVEC 2013 and AVEC 2014 were used to evaluate the proposed hybrid network combing LSTM with GIE and CNN with channel attention, and to compare with results of other published methods with comparable settings. This comparison is shown in Table 2, where all results are obtained only by speech.

Table 2 shows that using only the hand-craft short-term acoustic features [17, 18, 22] is difficult to mine the latent depression information. The performances of these conventional methods are much lower than those using deep learning methods [13, 23, 10, 11]. The method proposed by He et al. [13] considered only spatial information of the acoustic features and may lose important depression cues represented by the temporal information. In contrast, the hybrid network proposed by Niu et al. [10], which is the baseline of this paper, leveraged both the spatial and temporal information for depression detection, thus achieving a better performance. Zhao et al. [11] used self-attention networks trained on low-level acoustic features and DCNN trained on 3D Log-Mel spectrograms and tried to exploit the complementary segment-level features. The method proposed in this paper makes use of both LSTM with GIE and CNN with channel attention in the hybrid network, to improve the extracted spatial and temporal features for depression assessment, and obtains the best performance in RMSE on the test sets of both AVEC 2013 and AVEC 2014, which are 9.63 and 9.40, respectively. These results suggest the effectiveness of the proposed method for depression detection.

Table 2. Performance comparison between the proposed method and other published methods with comparable settings.

Network	AVEC 2013		AVEC 2014	
	RMSE	MAE	RMSE	MAE
Valster et al. [17, 18]	14.12	10.35	12.56	10.03
Meng et al. [22]	11.19	9.14	-	-
Jain et al. [23]	-	-	10.25	8.40
He et al. [13]	10.00	8.20	9.99	8.19
Niu et al. [10]	9.79	7.48	9.66	8.02
Zhao et al. [11]	9.65	**7.38**	9.57	7.94
Proposed	**9.63**	7.51	**9.40**	**7.37**

4. CONCLUSION

Long temporal information is crucial in emotion recognition, especially in depression assessment. To this end, this paper proposes a global information embedding module, which integrates the correlation information of the input sequence to the output sequence of the LSTM to enhance its discrimination. To further obtain the better temporal feature representation, this paper combines ℓ_p norm and LASSO in long-term features pooling from short-term features. Experiments on AVEC 2013 and AVEC 2014 verified the proposed method. Detailed comparisons with previous works also show the effectiveness of this method. The future work will focus on the joint representation of speech and its textual context in a latent space, by which to improve the depression level assessment by these two modalities.

5. ACKNOWLEDGEMENTS

This work is supported by the Open Project Program of the National Laboratory of Pattern Recognition (NLPR) (202200042, 202200012) and New Talent Project of Beijing University of Posts and Telecommunications (2021RC37).

6. REFERENCES

[1] Daniel Joseph France, Richard G Shiavi, Stephen Silverman, Marilyn Silverman, and M Wilkes, "Acoustical properties of speech as indicators of depression and suicidal risk," *IEEE Transactions on Biomedical Engineering*, vol. 47, no. 7, pp. 829–837, 2000.

[2] Nicholas Cummins, Julien Epps, Michael Breakspear, and Roland Goecke, "An investigation of depressed speech detection: Features and normalization," in *Twelfth Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2011, pp. 27–31.

[3] Jeffrey F Cohn, Tomas Simon Kruez, Iain Matthews, Ying Yang, Minh Hoai Nguyen, Margara Tejera Padilla, Feng Zhou, and Fernando De la Torre, "Detecting depression from facial actions and vocal prosody," in *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*. IEEE, 2009, pp. 1–7.

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

[5] Le Yang, Dongmei Jiang, Lang He, Ercheng Pei, Meshia Cédric Oveneke, and Hichem Sahli, "Decision tree based depression classification from audio video and language information," in *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, 2016, pp. 89–96.

[6] Nadee Seneviratne, James R Williamson, Adam C Lammert, Thomas F Quatieri, and Carol Y Espy-Wilson, "Extended study on the use of vocal tract variables to quantify neuromotor coordination in depression," in *INTERSPEECH*, 2020, pp. 4551–4555.

[7] Stefan Scherer, Giota Stratou, Jonathan Gratch, and Louis-Philippe Morency, "Investigating voice quality as a speaker-independent indicator of depression and ptsd," in *INTERSPEECH*, 2013, pp. 847–851.

[8] Andrea Carolina Trevino, Thomas Francis Quatieri, and Nicolas Malyska, "Phonologically-based biomarkers for major depressive disorder," *EURASIP Journal on Advances in Signal Processing*, vol. 2011, no. 1, pp. 1–18, 2011.

[9] Tuka Al Hanai, Mohammad M Ghassemi, and James R Glass, "Detecting depression with audio/text sequence modeling of interviews," in *INTERSPEECH*, 2018, pp. 1716–1720.

[10] Mingyue Niu, Jianhua Tao, Bin Liu, and Cunhang Fan, "Automatic depression level detection via lp-norm pooling," *Proc. INTERSPEECH, Graz, Austria*, pp. 4559–4563, 2019.

[11] Ziping Zhao, Qifei Li, Nicholas Cummins, Bin Liu, Haishuai Wang, Jianhua Tao, and Björn W Schuller, "Hybrid network feature extraction for depression assessment from speech," in *INTERSPEECH*, 2020, pp. 4956–4960.

[12] Le Yang, Hichem Sahli, Xiaohan Xia, Ercheng Pei, Meshia Cédric Oveneke, and Dongmei Jiang, "Hybrid depression classification and estimation from audio video and text information," in *Proceedings of the 7th Annual Workshop on Audio/visual Emotion Challenge*, 2017, pp. 45–51.

[13] Lang He and Cui Cao, "Automated depression analysis using convolutional neural networks from speech," *Journal of Biomedical Informatics*, vol. 83, pp. 103–111, 2018.

[14] Alice Othmani, Daoud Kadoch, Kamil Bentounes, Emna Rejaibi, Romain Alfred, and Abdenour Hadid, "Towards robust deep neural networks for affect and depression recognition from speech," *arXiv preprint arXiv:1911.00310v4*, 2020.

[15] Mohammed Senoussaoui, Milton Sarria-Paja, João F Santos, and Tiago H Falk, "Model fusion for multimodal depression classification and level detection," in *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*, 2014, pp. 57–63.

[16] Nadee Seneviratne and Carol Espy-Wilson, "Speech based depression severity level classification using a multi-stage dilated cnn-lstm model," *arXiv preprint arXiv:2104.04195*, 2021.

[17] Michel Valstar, Björn Schuller, Kirsty Smith, Florian Eyben, Bihan Jiang, Sanjay Bilakhia, Sebastian Schnieder, Roddy Cowie, and Maja Pantic, "Avec 2013: the continuous audio/visual emotion and depression recognition challenge," in *Proceedings of the 3rd ACM International Workshop on Audio/Visual Emotion Challenge*, 2013, pp. 3–10.

[18] Michel Valstar, Björn Schuller, Kirsty Smith, Timur Almaev, Florian Eyben, Jarek Krajewski, Roddy Cowie, and Maja Pantic, "Avec 2014: 3d dimensional affect and depression recognition challenge," in *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*, 2014, pp. 3–10.

[19] Takaya Taguchi, Hirokazu Tachikawa, Kiyotaka Nemoto, Masayuki Suzuki, Toru Nagano, Ryuki Tachibana, Masafumi Nishimura, and Tetsuaki Arai, "Major depressive disorder discrimination using vocal acoustic features," *Journal of Affective Disorders*, vol. 225, pp. 214–220, 2018.

[20] A McPherson and CR Martin, "A narrative review of the beck depression inventory (bdi) and implications for its use in an alcohol-dependent population," *Journal of Psychiatric and Mental Health Nursing*, vol. 17, no. 1, pp. 19–30, 2010.

[21] Jie Hu, Li Shen, and Gang Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7132–7141.

[22] Hongying Meng, Di Huang, Heng Wang, Hongyu Yang, Mohammed Ai-Shuraifi, and Yunhong Wang, "Depression recognition based on dynamic facial and vocal expression features using partial least square regression," in *Proceedings of the 3rd ACM International Workshop on Audio/Visual Emotion Challenge*, 2013, pp. 21–30.

[23] Varun Jain, James L Crowley, Anind K Dey, and Augustin Lux, "Depression estimation using audiovisual features and fisher vector encoding," in *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*, 2014, pp. 87–91.