INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

TDCA-Net: Time-Domain Channel Attention Network for Depression Detection

Cong Cai[1,2], *Mingyue Niu*[1,2], *Bin Liu*[1], *Jianhua Tao*[1,2,3], *Xuefei Liu*[1]

[1]National Laboratory of Pattern Recognition, Institute of Automation, CAS, Beijing, China
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
[3]CAS Center for Excellence in Brain Science and Intelligence Technology, Beijing, China
{cong.cai, mingyue.niu, liubin, jhtao, xuefei.liu}@nlpr.ia.ac.cn

Abstract

Depression is a psychiatric disorder and has many adverse effects on our society. Some studies have shown that speech signals are closely related to emotion and stress, and many speech-based automatic depression detection methods have been proposed. However, previous work is based on spectrogram or hand-crafted features, which may lose some useful information related to depression patterns. And there is no evidence that the filter bank designed from perceptual evidence is optimal for depression detection. In order to learn the more discriminative feature representation related to depression, we propose an end-to-end time-domain channel attention network (TDCA-Net) for depression detection. The TDCA-Net directly models time-domain speech signals based on dilated convolution block, which can increase the receptive field exponentially and aggregate multiscale contextual information associated with depression. Besides, we employ the efficient channel attention (ECA) module to model dependencies of channels and improve the sensitivity of the model to information related to depression. Experimental results on the AVEC2013 and the AVEC2014 datasets illustrate the effectiveness of our method.

Index Terms: end-to-end depression detection, time-domain, dilated convolution, channel attention, speech processing

1. Introduction

Depression is a psychiatric disorder, which has severely affected people's work and life [1, 2, 3]. Many studies have shown that speech signals have a close relationship with emotion and stress [4, 5, 6]. In recent years, to help clinicians to diagnose individual depression, many speech-based machine learning methods for depression detection have been proposed [7, 8, 9, 10, 11].

Many studies [12, 13, 14, 15] have found that depressive subjects are prone to possess a low dynamic range for the fundamental frequency, a slow rate of speech, and a relatively monotone delivery. Many works [7, 8, 16, 17] use hand-crafted features and spectrogram for depression detection. However, hand-crafted features and spectrogram may lose some useful information related to depression patterns. And it is unclear whether these features designed from perceptual evidence are the optimal for depression detection. An alternative method is to directly model the time-domain speech signals and extract features from raw waveform. In recent years, this approach has been successfully applied in many tasks, such as speech synthesis and speech separation [18, 19, 20, 21].

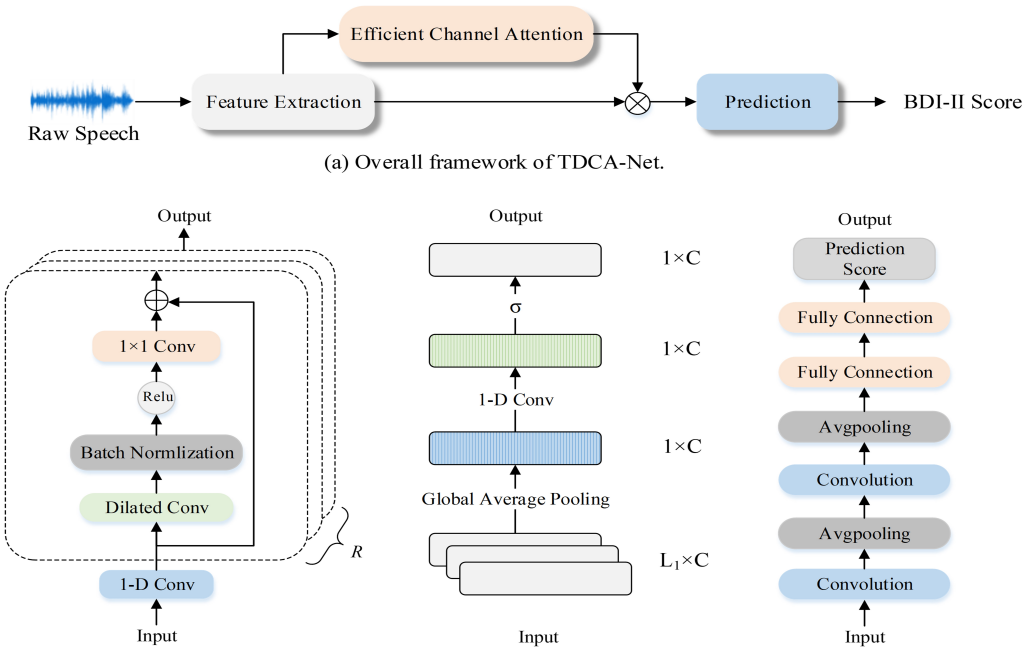
Some studies [16, 17, 22] have leveraged recurrent neural networks (RNNs) for depression analysis. And [11, 23, 24] have demonstrated the effectiveness of convolution neural networks (CNNs) for depression detection. However, RNNs is time-consuming because its calculation depends on the output of the previous time. The CNN's calculation is efficient, but it can't capture long-term dependence due to the limitation of convolution kernel size [25]. In this paper, we use the dilated convolution to extract the depression features from raw waveform, which can not only efficient in calculation, but also increase the receptive field exponentially. The dilated convolution has been used previously for emotion recognition [26] and depression detection [27]. Compared with hand-crafted features and spectrogram, directly using the raw waveform as input may lead to curse of dimensionality. The dilated convolution has large receptive field to avoid this problem effectively.

The attention mechanism has proven to improve the performance of CNNs [28, 29, 30]. Each channel has different contribution to depression detection. The channel attention mechanism can emphasize helpful feature channels related to depression. The SE-Net [29] presents an effective mechanism to learn channel attention and achieves promising performance. And the efficient channel attention (ECA) [28] module greatly reduces the number of parameters while maintaining high performance. In this paper, we take advantage of the ECA module to model the dependencies between channels, and improve model's representational power and sensitivity to depression information.

In this paper, we propose an end-to-end Time-Domain Channel Attention Network (TDCA-Net), which learns features fit for depression detection from raw waveform instead of spectrogram or hand-crafted features. The TDCA-Net consists of feature extraction module, efficient channel attention module and prediction module. In the feature extraction module, we propose the dilated convolution block to increase exponentially receptive field. And it also can aggregate multiscale contextual information associated with depression without losing resolution. Moreover, we explore the ECA [28] module to model dependencies between channels, which adaptively recalibrates channel feature responses and improves the representational power of the network. Finally, the prediction module predicts the individual's depression level score.

Our main contributions can be summarized as follows: 1) We propose dilated convolution blocks to directly model time-domain speech signals and extract depression features from the raw waveform, which can increase receptive field exponentially and aggregate multiscale contextual information related to depression. 2) We employ the ECA module to compute the weight of each feature channel by modelling dependencies between channels. Its local cross-channel interaction strategy achieves better performance with lower model complexity and improves the representational power of the network. 3) The experimental results on the AVEC2013 [9] and AVEC2014 [10] datasets demonstrate the effectiveness of our method.

http://dx.doi.org/10.21437/Interspeech.2021-1176



(a) Overall framework of TDCA-Net.

(b) Feature extraction module. (c) Efficient Channel attention module. (d) Prediction module.

Figure 1: *(a) Overall framework of the TDCA-Net. The TDCA-Net consists of the feature extraction module, the efficient channel attention module and the prediction module. (b) Illustration of the feature exctraction module. It contains a one-dimension convolution and R dilated convolution blocks. (c) Illustration of the ECA module. The ECA module generates a weight for each channel. (d) Illustration of the prediction module. It predicts the individual depression level score based on the previous features.*

2. Proposed Method

As shown in Figure 1 (a), the TDCA-Net consists of three modules: (1) The feature extraction module: It consists of a one-dimension convolution (1-D Conv) and R dilated convolution blocks. The raw waveform is input into a 1-D Conv, which downsamples to reduce the complexity of the model. And the dilated convolution blocks aggregate multiscale contextual information related to depression to extract discriminative features. (2) The efficient channel attention module: By modeling dependencies between channels, the ECA module calculates the weight of each feature channel. (3) The prediction module: It predicts the individual depression level score based on the weighted depression features.

2.1. Feature Extraction Module

The feature extraction module consists of a 1-D Conv and R dilated convolution blocks. The 1-D Conv transforms the input raw waveform into a C-dimensional representation, followed by a nonlinear activation function. The formula is as follows:

$$\mathcal{F}(\mathbf{X}) = \mathrm{ReLU}(\mathcal{H}(\mathbf{X})) \tag{1}$$

where $\mathbf{X} \in \mathbb{R}^{L_0 \times 1}$ is the input raw waveform, $\mathcal{H}(\mathbf{X}) \in \mathbb{R}^{L_1 \times C}$ denotes the output of the 1-D Conv, which contains C vectors with length L_1 each. And ReLU is a nonlinear activation function Rectified Linear Unit. The size of $\mathcal{F}(\cdot)$ output is $L_1 \times C$, and it is used as the input of the following R dilated convolution block.

2.1.1. Dilated Convolution

The ability of traditional convolution to process sequential signals is limited. Yu [25] proposed the dilated convolution, which can aggregate multiscale contextual information because of its exponentially growing receptive field. It can be defined as:

$$(F *_l k)(\mathbf{p}) = \sum_{\mathbf{s}+l\mathbf{t}=\mathbf{p}} F(\mathbf{s})k(\mathbf{t}) \tag{2}$$

where F is a discrete function, k is a discrete filter, $*_l$ is a dilated convolution or called l-dilated convolution. The filter of dilated convolution is applied over an area larger than its length by skipping input values with a certain step. It is equivalent to a convolution with a larger filter derived from the original filter by dilating it with zeros, but the dilated convolution is more efficient computationally.

In this paper, we use the dilated convolution to make the receptive field expand exponentially and the number of parameters only increase linearly. Depression is a continuous state rather than a instantaneous one, so dilated convolution is capable of aggregating contextual information related to depression at multiple scales.

2.1.2. Dilated Convolution Block

Motivated by the temporal convolution network (TCN) [19, 20, 31], we propose the dilated convolution block to extract features. Figure 1 (b) shows the structure of dilated convolution block. First the dilated convolution is used, and the batch normalization and the nonlinear activation function ReLU are followed. Then a 1×1 convolution (pointwise convolution) changes the number of output channels to make it consistent with the input channels. Finally we use the residual [32] path to speed up convergence and enable training of much deeper models. The output of each block is the input of the next block.

The dilation factor is doubled for every block up to a limit and then repeated: e.g. 1, 2, 4,..., 2^n,1, 2, 4,..., 2^n. The dilation factor increases exponentially to ensure to capture sufficiently large temporal contextual information related to depression. It expands the network's receptive field and captures the depres-

sion features of the entire speech with fewer layers of stacking. The size of the input and output are both $L_1 \times C$ in each block.

2.2. Efficient Channel Attention Module

The channel attention mechanism focuses on the relationship between channels, and it makes the model to automatically learn the importance of different feature channel. As illustrated in Figure 1 (c), we use the prevalent ECA [28] module in our method. It has proved that dimensionality reduction will have side effects on channel attention, and it is unnecessary to capture the relevance of all channels. In our method, the ECA module can focus on the learning of depression features by modeling channel dependencies and improve network's sensitivity to depression information.

Specifically, in order to take advantage of the global contextual information related to depression, the ECA module uses global average pooling independently for each channel to squeeze global information into a channel descriptor. And a fast 1-D Conv of size k is used to generate channel weights, which is designed to capture local cross-channel interaction by considering every channel and its k neighbors. Finally a Sigmoid function is used. The weights of the channels can be computed as:

$$\omega = \sigma(\mathcal{G}(\mathcal{L}(\mathbf{X}))) \tag{3}$$

where $\mathbf{X} \in \mathbb{R}^{L_1 \times C}$ is the input of ECA module as well as the output of the feature extraction module, L_1 and C are length and channel dimension. $\mathcal{L}(\mathbf{X}) = \frac{1}{L}\sum_{i=1}^{L} x_i$is channel global average pooling. $\mathcal{G}(\cdot)$ indicates 1-D Conv and σ is a Sigmoid function.

The ECA module effectively learns the importance related to depression of each channel and computes the weights. And it captures non-linear cross-channel interaction in an extremely lightweight way by avoiding channel dimensionality reduction. It effectively improves the performance of our convolutional architectures, only slightly increases the complexity and computational burden of the model. As illustrated in Figure 1 (c), the input size is $L_1 \times C$ and the output size is $1 \times C$.

2.3. Prediction Module

The input of the prediction module is the product of the output of the two previous modules and the size is $L_1 \times C$. The output is individual depression level score. Figure 1 (d) illustrates the structure of the prediction module, it consists of the convolution layers, average pooling and fully connection layers. The convolution layer is followed by a ReLU function, which increases the non-linear representation capable of the neural network. The average pooling and the fully connection layer not only aggregate depression information, but also control model complexity by reducing the dimensionality.

3. Experiments and Results

3.1. Dataset

In the AVEC2013 depression corpus, each subject needs to perform 14 different tasks according to the instructions on the computer screen. The length of each recording varies from 20 to 50 minutes, with an average duration of 25 minutes per recording. A total of 150 video clips from 84 subjects are divided into three parts: training, development and test sets. Each has 50 samples. There are 77 health samples and 73 depression samples.

The AVEC2014 depression corpus has two tasks named Northwind and FreeForm. In the Northwind the subjects are required to read excerpts from the German fable and in the FreeForm the subjects answer questions. The recording length of the Northwind task is between 31 s to 89 s, and that of the Freeform task is between 6 s and 248 s. In our experiment, we merge these two tasks into a new database. Namely, there are 100 samples in the training, development, and test sets respectively. And there are 77 health samples and 73 depression samples in each task. The level of depression in the AVEC 2013 and AVEC 2014 datasets is labelled with the Beck Depression Inventory-II (BDI-II) [33] score . The BDI-II score range from 0 to 63: 0–13 indicates no or minimal depression, 14–19 indicates mild depression, 20–28 indicates moderate depression, and 29–63 indicates severe depression.

In this paper, we conducted experiments on the AVEC2013 and AVEC2014 databases. For each database, the training set is used to train model. The development set is used to adjust the experimental parameters and verify the effectiveness of each module in the model. The test set is used to compare our method with other works.

3.2. Experimental Setup

First we use FFMPEG tool to extract audio from the video in the AVEC2013 and the AVEC2014 databases. Then each audio is divided into several segments, each segment is 3s and the overlap of two adjacent segments is 50%. The label of each segment is the BDI-II score of the corresponding audio. We sample the waveform at 8 kHZ for each segment, and generate the waveform points as input to the network. For the feature extraction module, there are 24 dilated convolution blocks in total and the highest dilation factor is 32. There are 64 kernels in the first 1-D convolution layer, last 1×1 convolutional layer and the efficient channel attention module. And there are 128 convolution kernels in the dilated convolution block. In the prediction module, the number of neurons in the two fully connected layers are 500 and 256. The optimizer is Adam, the batch size is 64, and the learning rate is 0.002. Finally, we predict the score of each segment, and calculate the average as the final score of the audio. The depression detection performance is assessed in terms of root mean square error (RMSE) and mean absolute error (MAE) between the prediction and reported BDI–II values. And our loss function is RMSE.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y_i})^2} \tag{4}$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y_i}| \tag{5}$$

3.3. Results and Discussion

3.3.1. Impact of Different Architectures

In this section, we compare the performance of different combination methods to evaluate the impact of each part on the development set of AVEC2013 and AVEC2014 databases. Table 1 presents the specific experimental results. It can be seen that the ECA module has a great contribution to depression detection, which emphasizes important feature channels by modelling dependencies between channels. In addition, the dilated convolution contributes more to improving the performance of depression detection. Because it not only increases the receptive field exponentially without dimensionality reduction, but also aggregates multiscale contextual information related to depression.

Table 1: *Performance comparison of different architectures on the development set of AVEC 2013 and AVEC 2014. The "TC" means that traditional convolution replace dilated convolution in the dilated convolution block. The "DC" indicates the dilated convolution, the "ECA" denotes the efficient channel attention module and the "P" refers to the prediction module.*

Metheds	AVEC2013		AVEC2014	
	RMSE	MAE	RMSE	MAE
TC+P	9.50	8.04	9.43	7.97
TC+ECA+P	9.22	7.84	9.17	7.78
DC+P	9.10	7.77	8.99	7.64
DC+ECA+P	**8.94**	**7.63**	**8.81**	**7.34**

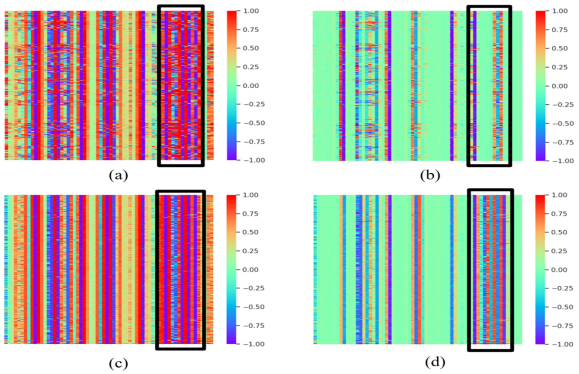


Figure 2: *Speech feature matrices of healthy and depressive individuals. The column is channel and the row is feature vector corresponding to the audio. (a) and (b) are the feature matrices output by feature extraction module and ECA module for the healthy subject of No. 316-1 in the AVEC2013 database. (c) and (d) are the results for the depressive subject of No. 246-1 in the AVEC2013 database.*

Based on our results, the best performance is achieved by the TDCA-Net model.

To further elaborate the effectiveness of the feature extraction module and the ECA module, We extract the feature vectors of the last layer of the two modules from individuals with different depression levels. The results are shown in Figure 2, these figures are drawn using the heatmap function in the Python. The horizontal axis indicates the channel and the vertical axis represents the feature vector. And the brighter the color, the larger the value. (a) and (b) are the feature matrices of the output of the feature extraction module and the ECA module from a healthy individual. And (c) and (d) are the results of a depressive individual. It can be found that the feature of healthy individual and depressive individual are significantly different. For example, the parts of (a) and (c) of enclosed by black boxes are more discriminative, which shows that the feature extraction module can automatically learn depression information effectively from speech.

3.3.2. Comparison with Previous Methods

In this section, we predict the depression level score on the test sets of AVEC2013 and AVEC2014 databases to compare with previous methods. Table 2 and Table 3 show these comparison results. They [7, 8, 9, 10, 17, 34, 35] use hand-crafted features as input to the model. And they [11, 16, 36] achieve good performance with spectrogram. However, the hand-crafted features and spectrogram may lose some useful information related to depression while extracting depression features. Only [11] also used the raw waveform as input to the network, but it uses the traditional convolution to process speech signals. The receptive field of traditional convolution is limited, so the final prediction result is inferior. Our approach not only learns discriminative feature representation related to depression from raw waveform, but also adaptively recalibrates channel feature responses by modelling dependencies between channels. It can be seen that our method has outstanding performance.

Table 2: *Performance comparison between the proposed model and other models on the test set of AVEC 2013.*

Metheds	RMSE	MAE
AVEC 2013 Audio Baseline [9]	14.12	10.35
PLS regression [34]	11.19	9.14
DCNN [11]	10.00	8.20
CNN-LSTM-SVR [17]	9.79	7.48
SAN-CNN [8]	9.65	7.38
STA-EEP [16]	9.50	7.14
Hierarchical Model [36]	**8.73**	7.32
TDCA-Net	9.22	**6.90**

Table 3: *Performance comparison between the proposed model and other models on the test set of AVEC 2014.*

Metheds	RMSE	MAE
AVEC 2014 Audio Baseline [10]	12.56	10.03
Fisher Vector Encoding [7]	11.51	9.74
PCA+Linear Regression [35]	10.28	8.07
DCNN [11]	9.99	8.19
CNN-LSTM-SVR [17]	9.66	8.02
SAN-CNN [8]	9.57	7.94
STA-EEP [16]	9.13	7.65
Hierarchical Model [36]	**8.82**	**6.80**
TDCA-Net	8.90	7.08

4. Conclusions

Spectrogram or hand-crafted features may lose some useful information related to depression. Motivated by this speculation, we propose an end-to-end time-domain channel attention network (TDCA-Net) for depression detection, which directly models the time-domain speech signals and extracts feature from raw waveform. The TDCA-Net uses the dilated convolution block to aggregate the contextual information associated with depression of the entire speech. And we employ the ECA module to emphasize the helpful channel related to depression by modelling dependencies between channels. Experimental results on AVEC2013 and AVEC2014 indicate that our approach achieves the promising performance on depression detection. In future, we will explore fusing text and video information to further improve the accuracy of depression detection.

5. Acknowledgements

This work is supported by the National Key Research & Development Plan of China (No.2017YFB1002804), the National Natural Science Foundation of China (NSFC) (No.61831022, No.61771472, No.61773379, No.61901473) and the Key Program of the Natural Science Foundation of Tianjin (Grant No. 18JCZDJC36300).

6. References

[1] C. D. Mathers and D. Loncar, "Projections of global mortality and burden of disease from 2002 to 2030," *PLoS medicine*, vol. 3, no. 11, p. e442, 2006.

[2] J. Olesen, A. Gustavsson, M. Svensson, H.-U. Wittchen, B. Jönsson, C. S. Group, and E. B. Council, "The economic cost of brain disorders in europe," *European journal of neurology*, vol. 19, no. 1, pp. 155–162, 2012.

[3] S. Saxena, M. Funk, and D. Chisholm, "Who's mental health action plan 2013-2020: what can psychiatrists do to facilitate its implementation?" *World Psychiatry*, vol. 13, no. 2, p. 107, 2014.

[4] D. R. Ladd, K. E. Silverman, F. Tolkmitt, G. Bergmann, and K. R. Scherer, "Evidence for the independent function of intonation contour type, voice quality, and f 0 range in signaling speaker affect," *The Journal of the Acoustical Society of America*, vol. 78, no. 2, pp. 435–444, 1985.

[5] K. R. Scherer, "Vocal affect expression: A review and a model for future research." *Psychological bulletin*, vol. 99, no. 2, p. 143, 1986.

[6] K. R. Scherer, R. Banse, H. G. Wallbott, and T. Goldbeck, "Vocal cues in emotion encoding and decoding," *Motivation and emotion*, vol. 15, no. 2, pp. 123–148, 1991.

[7] V. Jain, J. L. Crowley, A. K. Dey, and A. 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.

[8] Z. Zhao, Q. Li, N. Cummins, B. Liu, H. Wang, J. Tao, and B. W. Schuller, "Hybrid network feature extraction for depression assessment from speech," *Proc. Interspeech 2020*, pp. 4956–4960, 2020.

[9] M. Valstar, B. Schuller, K. Smith, F. Eyben, B. Jiang, S. Bilakhia, S. Schnieder, R. Cowie, and M. 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.

[10] M. Valstar, B. Schuller, K. Smith, T. Almaev, F. Eyben, J. Krajewski, R. Cowie, and M. 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.

[11] L. He and C. Cao, "Automated depression analysis using convolutional neural networks from speech," *Journal of biomedical informatics*, vol. 83, pp. 103–111, 2018.

[12] J. C. Mundt, P. J. Snyder, M. S. Cannizzaro, K. Chappie, and D. S. Geralts, "Voice acoustic measures of depression severity and treatment response collected via interactive voice response (ivr) technology," *Journal of neurolinguistics*, vol. 20, no. 1, pp. 50–64, 2007.

[13] L.-S. A. Low, N. C. Maddage, M. Lech, L. Sheeber, and N. Allen, "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*. IEEE, 2010, pp. 5154–5157.

[14] N. Cummins, J. Epps, M. Breakspear, and R. Goecke, "An investigation of depressed speech detection: Features and normalization," in *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[15] J. R. Williamson, T. F. Quatieri, B. S. Helfer, R. Horwitz, B. Yu, and D. D. Mehta, "Vocal biomarkers of depression based on motor incoordination," in *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*, 2013, pp. 41–48.

[16] M. Niu, J. Tao, B. Liu, J. Huang, and Z. Lian, "Multimodal spatiotemporal representation for automatic depression level detection," *IEEE Transactions on Affective Computing*, 2020.

[17] M. Niu, J. Tao, B. Liu, and C. Fan, "Automatic depression level detection via lp-norm pooling," *Proc. INTERSPEECH, Graz, Austria*, pp. 4559–4563, 2019.

[18] S. Pascual, A. Bonafonte, and J. Serra, "Segan: Speech enhancement generative adversarial network," *arXiv preprint arXiv:1703.09452*, 2017.

[19] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[20] Y. Luo and N. Mesgarani, "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 27, no. 8, pp. 1256–1266, 2019.

[21] S. Latif, R. Rana, S. Khalifa, R. Jurdak, and J. Epps, "Direct modelling of speech emotion from raw speech," *arXiv preprint arXiv:1904.03833*, 2019.

[22] Z. Zhao, Z. Bao, Z. Zhang, J. Deng, N. Cummins, H. Wang, J. Tao, and B. Schuller, "Automatic assessment of depression from speech via a hierarchical attention transfer network and attention autoencoders," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 2, pp. 423–434, 2019.

[23] L. Yang, H. Sahli, X. Xia, E. Pei, M. C. Oveneke, and D. 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.

[24] L. Yang, D. Jiang, X. Xia, E. Pei, M. C. Oveneke, and H. Sahli, "Multimodal measurement of depression using deep learning models," in *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*, 2017, pp. 53–59.

[25] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," *arXiv preprint arXiv:1511.07122*, 2015.

[26] H. Meng, T. Yan, F. Yuan, and H. Wei, "Speech emotion recognition from 3d log-mel spectrograms with deep learning network," *IEEE access*, vol. 7, pp. 125 868–125 881, 2019.

[27] Z. Huang, J. Epps, and D. Joachim, "Exploiting vocal tract coordination using dilated cnns for depression detection in naturalistic environments," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6549–6553.

[28] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "Eca-net: Efficient channel attention for deep convolutional neural networks," 2020.

[29] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[30] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.

[31] C. Lea, M. D. Flynn, R. Vidal, A. Reiter, and G. D. Hager, "Temporal convolutional networks for action segmentation and detection," in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 156–165.

[32] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[33] A. McPherson and C. 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.

[34] H. Meng, D. Huang, H. Wang, H. Yang, M. Ai-Shuraifi, and Y. 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.

[35] A. Jan, H. Meng, Y. F. B. A. Gaus, and F. Zhang, "Artificial intelligent system for automatic depression level analysis through visual and vocal expressions," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 10, no. 3, pp. 668–680, 2017.

[36] Y. Dong and X. Yang, "A hierarchical depression detection model based on vocal and emotional cues," *Neurocomputing*, vol. 441, pp. 279–290, 2021.