INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

FTA-net: A Frequency and Time Attention Network for Speech Depression Detection

Qifei Li, Dong Wang, Yiming Ren, Yingming Gao, Ya Li

School of Artificial Intelligence, Beijing University of Posts and Telecommunications, China

{liqifei, rym, yingming.gao, yli01}@bupt.edu.cn, dong1024mail@163.com

Abstract

Depression is one of the most common mental diseases nowadays, which seriously affects the health of individuals. Some researchers have shown an association between the level of depression and speech features in individuals, so a lot of automatic speech-based depression detection systems have been proposed. A number of studies utilized convolutional neural network (CNN) to realize the speech depression detection. However, most of these studies did not take into account that different frequencies and time steps in the speech spectrum features contribute unequally to the detection of depression. In order to extract more significant and distinctive features, this paper proposes an effective frequency-time attention (FTA) module for CNN, which is based on squeeze and excitation operations and can emphasize the time steps and frequencies associated with depression. Experimental results based on the AVEC 2013 and AVEC 2014 benchmarks demonstrate the effectiveness of our proposed method.

Index Terms: speech depression detection, frequency-time attention, residual network, convolutional neural network

1. Introduction

Depression is a typical psychological disorder. and the clinical symptoms are significant and persistent depression mood, loss of interest, lack of vigour, etc. As the condition deteriorates, it may even cause patients to suicide and self-mutilation [1]. Early detection and diagnosis are crucial for depression prevention and treatment. Therefore, there is an urgent need to develop an automatic depression detection method that is low-cost, efficient and universal.

Several studies have shown that speech features are significantly related to the severity of depression and can be used to distinguish depressed and non-depressed individuals [2, 3, 4]. Because of the above studies, more and more researchers implemented automatic depression detection systems based on speech signal [5, 6, 7, 8, 9, 10, 11]. In recent years, long-short-term memory network (LSTM) and CNN have become the most commonly used network structures for depression detection. Ma et al. [8] proposed a DepAudioNet consisting of LSTM and CNN to assess the depression level. He et al. [9] used multi-branch CNN as feature extractor to extract depression-related characteristics from Mel-frequency cepstral coefficient (MFCC) and spectrogram. Niu et al. [10] proposed a hybrid network composed of LSTM and CNN to extract features from short-term MFCC segments simultaneously. EmoAudioNet [11] was a two-branch CNN model for extracting deep features associated with depression from MFCC and spectrogram respectively. Li et al. [12] took advantage of channel attention and global information embedding to improve the feature capturing ability of CNN and LSTM respectively, thus improving the performance of depression detection.

The spectrogram used for the speech task is different from the image used for traditional image recognition. Both dimensions of a picture represent pixel coordinates, but the subsidiary information carried in the time and frequency dimensions of the speech spectrogram is not similar. Some researchers consider the above problem according to their speech task needs, such as language identification, acoustic event detection, speech emotion recognition, speech enhancement, etc. [13, 14, 15, 16, 17], and design different time-frequency attention mechanisms to focus on the information in both dimensions of the spectrogram. The information such as gender and speech speed characterized by different frequencies and time steps are also very important for depression recognition [18, 19, 20]. Therefore, paying attention to the information subsidiary in the time-frequency dimension is crucial for depression detection. However, very few depression detection studies pay attention to this issue. The above time-frequency attention methods are used for sequence models and not applicable to 2D CNN, which have been shown to be effective for the detection of speech depression [8, 9, 10, 11, 12]. In addition, amplitude and phase spectra are obtained during the extraction of spectrograms, and most of all research works use only the amplitude information as input of the model. However, the phase information is important for speech quality and intelligibility [21, 22]. Hence, it may also imply depression-related information, which needs to be verified.

Therefore, we propose an end-to-end frequency-time attention residual network (FTA-net) which contains multiple FTAs. The FTA can help 2D CNN emphasize the contribution of different frequencies and time steps in speech depression detection to improve the detection performance by making full use of the subsidiary information in the time-frequency dimension. The input of the model is complex spectrogram, which consists of real and imaginary spectrogram, thus it contains not only amplitude but also phase information of speech signal. We make FTA-net fit this feature so that it learns all the information in the speech signal as much as possible to predict the level of depression. In addition, inspired by [23, 24], we use the attentive statistics pooling (ASP) to aggregate the output of FTA-net which first focuses on depression-related frames using an attention mechanism and then aggregates the frame-level features into utterance-level representations for the final depression level assessment.

The main contributions of this study can be summarized as follows: i) We propose an end-to end frequency-time residual network with attentive statistics pooling for speech depression detection; ii) We validate the effectiveness of phase information on speech depression detection through ablation experiments.

10.21437/Interspeech.2023-296

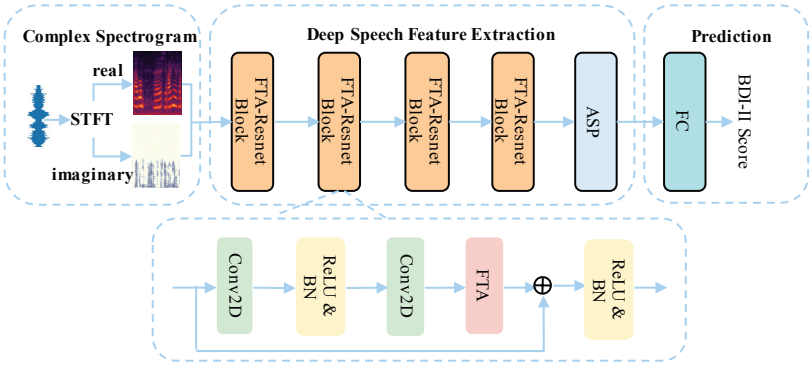

Figure 1: *Framework of the proposed FTA-net. BN and FC denote the batch normalization and the fully connected layer respectively.*

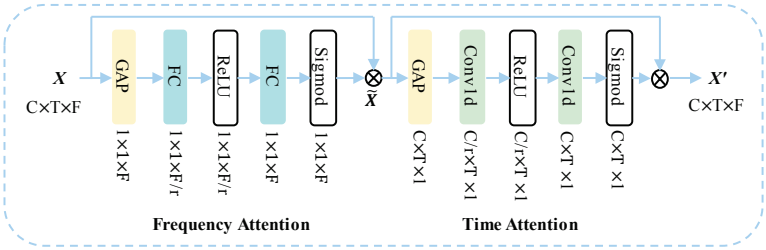

Figure 2: *Framework of the proposed FTA. C,T and F denote the dimension of channel, time and frequency respectively. GAP and FC represent global average pooling and fully connected layer respectively. r is reduction ratio.*

2. Proposed Method

Fig. 1 illustrates the overall framework of FTA-net which consists of three parts. The first is the complex spectrogram which is the input of the model. The second is the FTA-Resnet block which is composed by FTA and CNN. FTA can help model to capture the depression-related information in the dimension of frequency and time. The third part is ASP module, in which attention mechanism is applied to aggregate deep frame-level features into utterance-level features for depression severity prediction. Section 2.1 and section 2.2 describe the FTA and ASP modules in detail, respectively.

2.1. Frequency-time Attention

Squeeze-and-excitation block (SE block) is a well-known and very effective channel attention for CNN [25]. It can automatically learn the importance of each channel in the feature map by *squeeze* and *excitation* operations. These operations emphasize the important channels and suppress the information that are not relevant to the task, thus improving the performance of the model. Therefore, we borrow its idea of *squeeze* and *excitation* to implement FTA. A diagram illustrating the structure of a FTA module is shown in Fig. 2.

2.1.1. Frequncy Attention

According to Fig. 2, it firstly uses global average pooling to *squeeze* the input $X \in R^{C\times T\times F}$ into frequency-wise statistical embedding $\mathbf{z} \in R^{1\times 1\times F}$. The f-th element of $\mathbf{z}$ is calculated as follows:

$$z_f = \mathbf{F}_{sq}(\mathbf{x}_f) = \frac{1}{C\times T}\sum_{i=1}^{C}\sum_{j=1}^{T} x_f(i,j) \qquad (1)$$

where $\mathbf{F}_{sq}$ means the *squeeze* operation. The $\mathbf{x}_f$ is the f-th 2D channel-time matrix of input $X = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_f]$ which is the output of CNN.

Then we employ the *excitation* operation to take advantage of the statistical embedding aggregated in the *squeeze* operation and fully capture frequency-wise dependencies. The *excitation* operation consists of two fully connected layers (FC) as shown in Equation 2:

$$\mathbf{s} = \mathbf{F}_{ex}(\mathbf{z}, \mathbf{W}) = Sigmoid(\mathbf{W}_2 ReLU(\mathbf{W}_1\mathbf{z})) \qquad (2)$$

where $\mathbf{F}_{ex}$ represents the *excitation* operation. The $\mathbf{W}_1 \in R^{\frac{F}{r}\times F}$ and $\mathbf{W}_2 \in R^{F\times\frac{F}{r}}$ are learnable weight matrix of the two FC layers. The parameter r means the reduction factor in order to reduce model complexity and improve generalization ability. $\mathbf{s}$ means the attention weight of every frequency.

Finally, the output $\tilde{X} \in R^{C\times T\times F}$ of frequency attention (FA) can be obtained by multiplying $\mathbf{s}$ and X:

$$\tilde{X} = \mathbf{s}X \qquad (3)$$

2.1.2. Time Attention

Here, we also use *squeeze* and *excitation* operations to achieve time attention (TA). Since we not only want to calculate the attention weights of the T dimension but also want to use the channel information to fuse the results of the frequency and time attention, we only *squeeze* the F dimension and use 1D convolution to implement the *excitation* operation. 1D convolution of the channel dimension can be considered as the channel attention weights to fuse the results of the frequency and time attention. The *squeeze* operation is calculated as shown in the Equation (4):

$$y_t = \mathbf{F}_{sq}(\tilde{\mathbf{x}}_t) = \frac{1}{F}\sum_{i=1}^{F}\tilde{x}_t(i) \qquad (4)$$

where $\tilde{X} = [\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_t]$ is output of FA, and $\mathbf{y} \in R^{C\times T\times 1} = [y_1, y_2, \ldots, y_t]$ means the time and channel statistical embedding by squeezing F dimension.

The *excitation* operation is calculated as Equation (5). Where $\mathbf{W}_3 \in R^{\frac{C}{r}\times C}$ and $\mathbf{W}_4 \in R^{C\times\frac{C}{r}}$ are kernel matrix of 1D CNN. r are also the reduction factor.

$$\mathbf{u} = \mathbf{F}_{ex}(\mathbf{y}, \mathbf{W}) = Sigmoid(\mathbf{W}_4 ReLU(\mathbf{W}_3\mathbf{y})) \quad (5)$$

Finally, the output X' of FTA can be expressed as follow:

$$X' = \mathbf{u}\tilde{X} \quad (6)$$

2.2. Attentive Statistics Pooling

ASP first rescales the frame-level features using the attention mechanism, then calculates the mean and variance of the rescaled features for dimensionality reduction, and realizes the feature transformation from frame-level to utterance-level. ASP is implemented as follows:

$$\alpha_i = Softmax(tanh(\mathbf{W}_5 h_i)) \quad (7)$$

$$V = Concate(\mu(\alpha_i x_i'), \sigma(\alpha_i x_i')) \quad (8)$$

where α_t is attention weight for each frame-level feature. μ and σ denote the mean and variance of features which are rescaled by attention mechanism. Finally, the utterance-level statistical features V are fed to FC layer for regression.

3. Experiments and Results

3.1. Experimental Corpus

The AVEC2013 benchmark has 84 subjects, ranging in age from 18 to 63 years old, with an average age of 31. There are 150 video recordings in benchmark. The duration of the recordings varies from 20 to 50 minutes, and 25 minutes on average. The benchmark is divided into 3 parts, training, validation and test set, of which each contains 50 video samples.

The AVEC2014 is a subset of AVEC2013 and contains two tasks, NorthWind and FreeForm. Each task contains 150 video samples, and the training, validation and test sets include 50 samples respectively. The duration of videos in the NorthWind is distributed from 31 to 89 seconds. and that of the videos in FreeForm is between 6 to 248 seconds. In the experiments of this paper, we combine the samples of both tasks, in other words, the training, validation and test sets each contain 100 samples. Both benchmarks are labeled with scores from the Beck Depression Inventory-II (BDI-II), a standard depression self-report inventory [26].

3.2. Experimental Setup

First of all, we transcribe the video files in those benchmarks into audio files and resample the audio to 8 kHz. Due to the sample size limitation of the benchmarks, we use a data augmentation approach using a sliding window with a size of 3 seconds and 50% overlap to divide the raw audio into multiple speech segments. For the STFT, the hanning window is 50 ms with a shift of 12.5 ms, and the fast Fourier transform (FFT) points are 512. Before the complex spectrogram is fed into FTA-Resnet block, it first goes into a 2D CNN with a kernel size of 9×3 and a stride of (3, 1). The number of convolution kernels in the four FTA-Resnet blocks is 64, 128, 256, and 512, respectively, and the size of the kernel is 3. The stride of first CNN in the last three blocks is 2. The rest of the parameters in FTA-net are as same as Resnet18 [27]. The number of neurons in the FC layer is 256. All parameters in the model are about 15M. The GPU used is RTX3090, the optimizer is Adam, the learning rate is 0.002, and the size of batch is 64. During training, the training set needs class re-balancing. In the end, the average of prediction results of all speech segments of the subject is regarded as final depression score. The results of AVEC 2014 are obtained by fine-tuning the model from AVEC 2013. The root mean square error (RMSE) and the mean absolute error (MAE) are used to evaluate the performance of our proposed method.

Table 1: *Performance comparison of different attention and input on the test set of AVEC 2013 and AVEC 2014.*

Method	AVEC 2013		AVEC 2014	
	RMSE	MAE	RMSE	MAE
Resnet	10.39	8.22	10.35	8.06
F-net	10.26	8.16	10.27	7.91
T-net/o	10.30	8.07	10.29	7.88
T-net	10.03	7.67	9.96	7.65
SE-net	9.77	7.52	9.81	7.53
TFA-net	10.13	7.55	10.12	7.60
FTA-net-R	9.85	7.69	9.91	7.65
FTA-net	**9.58**	**7.42**	**9.60**	**7.31**

3.3. Results and Discussion

3.3.1. Ablation tests and analysis

Table 1 shows the results of the ablation experiments on AVEC 2013 and AVEC 2014 with different attention mechanisms to demonstrate the effectiveness of the proposed method. Resnet indicates that the CNN blocks do not use any attention. T-net, F-net denote Resnet with TA and FA respectively. T-net/o indicates the same structure as F-net, without channel information. TFA-net means that TA is used first in Resnet block, and then FA. FTA-net-R represents that only the real spectrogram is used as input to FTA-net.

As shown in Table 1, we can observe that proposed FTA-net consisitently achieves the best performance on both benchmarks. Moreover, T-net and F-net outperform Resnet in both RMSE and MAE, indicating that subsidiary information in the time and frequency dimensions benefit to speech depression detection and can exploit the attention mechanism to emphasize the depression-related time steps and frequencies. F-net and T-net/o have similar performance but are worse than T-net. After analysis, this is because the channel information is included in the process of *squeeze* and *excitation* in TA, which also explains why FTA-net performs better than TFA-net. In addition, we also compare the effects of real and complex spectrogram as model inputs on depression detection, and the experimental results show that the performance of complex spectrogram outperform real spectrogram, which demonstrates that not only the amplitude information but also phase information plays an important part in speech depression detection.

To better demonstrate the contribution of time and frequency attention, we plot the 40th channel in the output of the third attention for the different models, as shown in Fig. 3. The brighter the bar, the more important it is for depression detection. As we can see from the Fig. 3, the different time steps

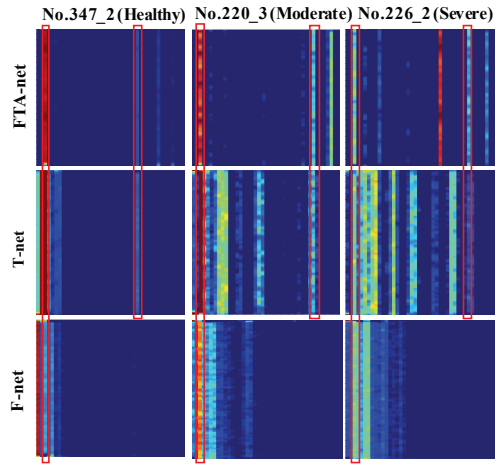

Figure 3: *Attention heatmap of subjects with different depression levels drawn from the time and frequency attention modules. 347_2, 220_3, 226_2 are the numbers of subjects with different levels of depression. The column of subplot is frequency and the row of subplot is time step.*

and frequencies do contribute differently. The three attention methods share a common focus on time steps and frequencies in red color, demonstrating that these steps and frequencies are particularly important for depression detection. The remaining differences in the figure determine their performance. The feature maps of the different depression levels in the FTA-net are significantly discriminatory, whereas the feature maps of moderate and severe depression in the T-net and the F-net have high similarity, indicating that they are difficult to distinguish moderate and severe depressed subjects.

For a more precise illustration, we calculate the Euclidean distance between the above feature maps. From the direction of the rows of Table 2, the maximum Euclidean distance is obtained for each group, indicating that the FTA can help the model to distinguish subjects with different depression levels. However, the Euclidean distances of H-M and H-S for T-net and F-net are similar, suggesting that they can not distinguish subjects with moderate and severe depression. The same is evidenced by their M-S values. The above analysis is consistent with the results in Table 1, which further suggests the effectiveness of FTA-net.

Table 2: *Euclidean distances between feature maps for different depression levels in different models. H, M and S denote healthy, moderate and severe depression subject groups, respectively.*

Methods	H-M	H-S	M-S
FTA-net	41.96	55.89	26.19
T-net	17.93	18.36	7.66
F-net	3.00	3.29	1.04

3.3.2. Performance comparison with previous methods

In this subsection, we highlight the effectiveness of our proposed method by comparing it with key results from previous studies on AVEC 2013 and AVEC 2014 benchmarks. The comparison results are shown in Table 3 and Table 4. It can be observed that the best RMSE (9.58) on the AVEC 2013 test set and the best MAE (7.31) on AVEC 2014 test set were obtained by our FTA-net. These results prove the effectiveness of our proposed method. In addition, FTA-net is an end-to-end approach that is more convenient for practical application than the above studies [10, 12, 28] with better performance.

Table 3: *Performance comparison between the proposed model and other models on the test set of AVEC 2013 benchmark.*

Methods	RMSE	MAE
AVEC 2013 Audio Baseline [5]	12.56	10.03
PLS regression [7]	10.25	8.40
DCNN [9]	9.99	8.19
CNN-LSTM-SVR [10]	9.66	8.02
SAN-CNN-SVR [28]	9.65	**7.38**
GIE-LSTM-CNN-SVR [12]	9.63	7.51
FTA-net (ours)	**9.58**	7.42

Table 4: *Performance comparison between the proposed model and other models on the test set of AVEC 2014 benchmark.*

Methods	RMSE	MAE
AVEC 2014 Audio Baseline [6]	12.56	10.03
Fisher Vector Encodeing [29]	10.25	8.40
DCNN [9]	9.99	8.19
CNN-LSTM-SVR [10]	9.66	8.02
SAN-CNN-SVR [28]	9.57	7.97
GIE-LSTM-CNN-SVR [12]	**9.40**	7.37
FTA-net (ours)	9.60	**7.31**

4. Conclusions

In this work, considering that different time steps and frequencies on the spectrogram contribute unequally to the speech depression detection and that some subsidiary information in both dimensions of the spectrogram contributes to speech depression detection, we proposed FTA-net with input of complex spectrogram that can focus on both time and frequency domains to improve the performance of depression detection. In this paper, we used a large number of ablation experiments to verify the effectiveness of our proposed method, and the results of ablation experiments on AVEC 2013 and AVEC 2014 benchmarks suggest that our proposed FTA module can effectively focus on the depression-related time steps and frequencies in the features. In addition, we also verified that phase information is helpful for depression detection. In future work, considering the long-term context-dependent nature of depression detection, we will explore the variants of FTA that can be used in temporal networks to implement speech depression detection.

5. Acknowledgements

This work was supported by the the National Natural Science Foundation of China (NSFC) (No. 62271083), Open Project Program of the National Laboratory of Pattern Recognition (NLPR) (No. 202200042, No. 202200012) and the Fundamental Research Funds for the Central Universities (No. 2023RC13).

6. References

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

[2] M. Cannizzaro, B. Harel, N. Reilly, P. Chappell, and P. J. Snyder, "Voice acoustical measurement of the severity of major depression," *Brain and Cognition*, vol. 56, no. 1, pp. 30–35, 2004.

[3] J. C. Mundt, A. P. Vogel, D. E. Feltner, and W. R. Lenderking, "Vocal acoustic biomarkers of depression severity and treatment response," *Biological Psychiatry*, vol. 72, no. 7, pp. 580–587, 2012.

[4] A. S. Cohen, Y. Kim, and G. M. Najolia, "Psychiatric symptom versus neurocognitive correlates of diminished expressivity in schizophrenia and mood disorders," *Schizophrenia Research*, vol. 146, no. 1-3, pp. 249–253, 2013.

[5] 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.

[6] 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.

[7] 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.

[8] X. Ma, H. Yang, Q. Chen, D. Huang, and Y. Wang, "Depaudionet: An efficient deep model for audio based depression classification," in *Proceedings of the 6th international workshop on audio/visual emotion challenge*, 2016, pp. 35–42.

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

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

[11] A. Othmani, D. Kadoch, K. Bentounes, E. Rejaibi, R. Alfred, and A. Hadid, "Towards robust deep neural networks for affect and depression recognition from speech," in *Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10–15, 2021, Proceedings, Part II*. Springer, 2021, pp. 5–19.

[12] Y. Li, M. Niu, Z. Zhao, and J. Tao, "Automatic depression level assessment from speech by long-term global information embedding," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 8507–8511.

[13] X. Miao, I. McLoughlin, and Y. Yan, "A New Time-Frequency Attention Mechanism for TDNN and CNN-LSTM-TDNN, with Application to Language Identification," in *Proc. Interspeech 2019*, 2019, pp. 4080–4084.

[14] J. Zhang, W. Ding, J. Kang, and L. He, "Multi-Scale Time-Frequency Attention for Acoustic Event Detection," in *Proc. Interspeech 2019*, 2019, pp. 3855–3859.

[15] B. Wu and X.-P. Zhang, "Environmental sound classification via time–frequency attention and framewise self-attention-based deep neural networks," *IEEE Internet of Things Journal*, vol. 9, no. 5, pp. 3416–3428, 2021.

[16] K. Liu, C. Wang, J. Chen, and J. Feng, "Time-frequency attention for speech emotion recognition with squeeze-and-excitation blocks," in *MultiMedia Modeling: 28th International Conference, MMM 2022, Phu Quoc, Vietnam, June 6–10, 2022, Proceedings, Part I*. Springer, 2022, pp. 533–543.

[17] Q. Zhang, X. Qian, Z. Ni, A. Nicolson, E. Ambikairajah, and H. Li, "A time-frequency attention module for neural speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 462–475, 2022.

[18] T. Taguchi, H. Tachikawa, K. Nemoto, M. Suzuki, T. Nagano, R. Tachibana, M. Nishimura, and T. Arai, "Major depressive disorder discrimination using vocal acoustic features," *Journal of Affective Disorders*, vol. 225, pp. 214–220, 2018.

[19] E. Toto, M. Tlachac, and E. A. Rundensteiner, "Audibert: A deep transfer learning multimodal classification framework for depression screening," in *Proceedings of the 30th ACM international conference on information & knowledge management*, 2021, pp. 4145–4154.

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

[21] K. Paliwal, K. Wójcicki, and B. Shannon, "The importance of phase in speech enhancement," *Speech Communication*, vol. 53, no. 4, pp. 465–494, 2011.

[22] C. Fan, J. Tao, B. Liu, J. Yi, Z. Wen, and X. Liu, "End-to-end post-filter for speech separation with deep attention fusion features," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1303–1314, 2020.

[23] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification," in *Proc. Interspeech 2020*, 2020, pp. 3830–3834.

[24] D. Wang, Y. Ding, Q. Zhao, P. Yang, S. Tan, and Y. Li, "ECAPA-TDNN Based Depression Detection from Clinical Speech," in *Proc. Interspeech 2022*, 2022, pp. 3333–3337.

[25] 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.

[26] A. T. Beck, R. A. Steer, and G. Brown, "Beck depression inventory–ii," *Psychological Assessment*, 1996.

[27] 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.

[28] 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," in *Proc. Interspeech 2020*, 2020, pp. 4956–4960.

[29] 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.